DOCUMENT RESUME

ED 471 665 TM 034 694


AUTHOR Lee, Won-Chan; Brennan, Robert L.; Kolen, Michael J.


TITLE Interval Estimation for True Scores under Various Scale Transformations. ACT Research Report Series.


INSTITUTION American Coll. Testing Program, Iowa City, IA.
REPORT NO ACT-RR-2002-5
PUB DATE 2002-11-00
NOTE 78p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC04 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Error of Measurement; *Estimation (Mathematics); Scaling; Simulation; *True Scores
IDENTIFIERS *Confidence Intervals (Statistics); *Nonlinear Transformations



ABSTRACT

This paper reviews various procedures for constructing an interval for an individual's true score given the assumption that errors of measurement are distributed as binomial. This paper also presents two general interval estimation procedures (i.e., normal approximation and endpoints conversion methods) for an individual's true scale score; compares the various interval estimation procedures through computer simulation studies by evaluating how close actual coverage probabilities are to selected nominal levels (i.e., 0.95, 0.68, and 0.50); and provides some practical guidelines for the use of the interval estimation procedures. To examine the effects of different types of scale scores, four nonlinearly transformed scale scores are used. The conditional confidence intervals using conditional standard errors of measurement are recommended over the traditional confidence intervals using the overall standard error of measurement, especially for lower nominal levels. The score confidence interval, Bayes confidence interval, and credibility interval tend to provide the actual coverage probabilities that are closest to the nominal levels, on average. Results for scale score intervals appear to favor the endpoints conversion method using the true-score conversions over the normal approximation approach. (Contains 7 tables, 27 figures, and 46 references.) (Author/SLD)

ACT Research Report Series 2002-5



# Interval Estimation for True Scores Under Various Scale Transformations

Won-Chan Lee

Robert L. Brennan

Michael J. Kolen











November 2002

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

# Interval Estimation for True Scores Under Various Scale Transformations

**Won-Chan Lee**
ACT, Inc.

**Robert L. Brennan**
**Michael J. Kolen**
The University of Iowa

## Table of Contents

## Abstract

This paper reviews various procedures for constructing an interval for an individual's true score given the assumption that errors of measurement are distributed as binomial. This paper also presents two general interval estimation procedures (i.e., normal approximation and endpoints conversion methods) for an individual's true scale score; compares the various interval estimation procedures through computer simulation studies by evaluating how close actual coverage probabilities are to selected nominal levels (i.e., .95, .68. and .5); and provides some practical guidelines for use of the interval estimation procedures. To examine the effects of different types of scale scores, four non-linearly transformed scale scores are employed. The conditional confidence intervals using conditional standard errors of measurement are recommended over the traditional confidence intervals using the overall standard error of measurement, especially for lower nominal levels. The score confidence interval, Bayes confidence interval, and credibility interval tend to provide the actual coverage probabilities that are closest to the nominal levels, on average. Results for scale score intervals appear to favor the endpoints conversion method using the true-score conversions over the normal approximation approach.

## Interval Estimation for True Scores Under Various Scale Transformations[1]

### Introduction

One of the goals of educational and psychological measurement is to estimate examinees' true scores. A point estimate of the true score may not be very meaningful without being accompanied by some measure of the errors involved in a measurement procedure. Standard errors of measurement (SEMs) typically are used to report the amount of measurement error in test scores. One very practical use of SEMs is in making inferences about an examinee's true score via confidence intervals (Lord & Novick, 1968). Traditionally, confidence intervals have been constructed using a strong assumption that measurement errors are normally distributed and the standard error of measurement is the same for all examinees (Feldt & Brennan, 1989). The traditional definition of SEM (i.e., same for all examinees) is sometimes called the overall SEM in the sense that it is an average SEM for all examinees in the population.

A large volume of measurement literature, however, has been devoted to the theoretical developments and empirical justification for SEMs that differ at different points on the score scale (Brennan, 1996, 1998; Feldt, 1984; Feldt & Qualls, 1996; Feldt, Steffen, & Gupta, 1985; Lord, 1955, 1957, 1984; Mollenkopf, 1949; Qualls-Payne, 1992; Thorndike, 1951). As opposed to the overall SEM, the SEMs associated with individuals' specific score levels are referred to as conditional SEMs. When a confidence interval is constructed for an examinee with a particular true score using the examinee's conditional SEM, the interval is referred to here as a conditional confidence interval. Note that we can form a confidence interval for an isolated individual using either the overall SEM or conditional SEM. It has been suggested in the literature, however, that

[1] A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, April 1999. The authors thank Chiou-Yueh Shyu and Matthew Schulz for their helpful comments on the paper.

confidence intervals be based on conditional SEMs, not on the SEMs for a test as a whole (Feldt et al., 1985; Harvill, 1991).

The SEMs and confidence intervals can be stated in terms of both raw scores (i.e., number-correct scores) and transformed scale scores. In recent years, some procedures have been developed for estimating conditional SEMs for scale scores (Brennan & Lee, 1997, 1999; Feldt & Qualls, 1998; Kolen, Hanson, & Brennan, 1992; Kolen & Wang, 1998; Kolen, Zeng, & Hanson, 1996; Wang, Kolen, & Harris, 2000). These procedures could be readily used to construct a conditional confidence interval for an individual's true scale score. The conditional confidence intervals for raw scores or scale scores using conditional SEMs have not been considered extensively.

Bayesian inference also provides a means of constructing an interval for an individual. The resultant intervals are often called credibility intervals (Novick & Jackson, 1974, pp. 119-126). A credibility interval for an examinee gives information about the distribution of the examinee's true score (i.e., posterior distribution), given one's prior knowledge (i.e., prior distribution) and the observed score. As discussed later, credibility intervals differ from confidence intervals in several ways. These two general approaches are compared in this paper in terms of estimation accuracy.

The present paper (1) reviews various procedures for constructing intervals for raw scores, (2) presents two general interval estimation procedures for scale scores, (3) compares the various interval estimation procedures through computer simulation studies by evaluating how close the actual coverage probabilities are to the nominal levels, and (4) provides some practical guidelines for use of the interval estimation procedures. To examine the effects of raw-to-scale score transformations, four different types of scale scores are used: developmental standard scores (DSSs), grade equivalents (GEs), percentile ranks (PRs), and stanines (STs), which are all non-linear transformations of raw scores. Developmental standard scores are the primary score scale that is reported to test users for the *Iowa Tests of Basic Skills* (ITBS) (Hoover, Hieronymus, Frisbie, &

Dunbar, 1993a). Petersen, Kolen, and Hoover (1989) describe these four types of scale scores in some detail.

Note that some interval estimation procedures discussed in this paper were developed only for the binomial parameter. Thus, to establish comparability across procedures, the binomial error model is considered to be an underlying distribution of errors for all procedures. Accordingly, the simulation is based on the binomial error model as well.

## Intervals for Raw Scores

This paper considers six different interval procedures for raw scores: (1) conditional confidence intervals using conditional SEMs, (2) traditional overall confidence intervals using the overall SEM, (3) score confidence intervals, (4) Bayes confidence intervals, (5) Clopper-Pearson exact confidence intervals, and (6) credibility intervals. The first four procedures, in effect, are based on normal distribution assumptions in one way or another, while the Clopper-Pearson exact confidence interval uses the binomial distribution, which is the "exact" distribution for the observed scores for a person. Credibility intervals often use a beta distribution to describe the posterior distribution. The binomial error model and some issues related to confidence intervals for the binomial parameter are discussed first followed by the overviews of the interval procedures.

### *Confidence Intervals for a Binomial Parameter*

Let $X$ denote a random variable for an examinee's observed number-correct score. Further, let $\tau$ be the true score for an examinee, which is defined as the expected value of the observed scores obtained from repeated measurements. Under the binomial error model, the conditional distribution of observed score $X$ given an individual's proportion-

correct true score, $\pi = \tau / k$, on a test consisting of $k$ dichotomously-scored items is the binomial distribution (Lord & Novick, 1968):

$$\Pr(X = x \mid \pi) = \binom{k}{x} \pi^x (1-\pi)^{k-x}. \tag{1}$$

That is, $X$ is binomially distributed with a mean of $k\pi = \tau$ and a standard deviation of $\sqrt{k\pi(1-\pi)}$.

Let $\bar{X}$ be a random variable for the observed proportion-correct score, and consider the problem of determining a confidence interval for $\pi$. From the Central Limit Theorem, $(\bar{X} - \pi)/\sqrt{\pi(1-\pi)/k}$ has a limiting standard normal distribution, $N(0,1)$. Using theorems on limiting distributions (Hogg & Craig, 1995, pp. 253-255), it can be shown that $(\bar{X} - \pi)/\sqrt{\bar{X}(1-\bar{X})/k}$ has a limiting distribution of $N(0,1)$ as well. Thus, we have

$$\Pr\left[-z_\gamma \le \frac{\bar{X} - \pi}{\sqrt{\bar{X}(1-\bar{X})/k}} \le z_\gamma\right] \approx \gamma, \tag{2}$$

where $\gamma$ is a probability value, and $z_\gamma$ denotes $(1+\gamma)/2$ th quantile of the standard normal distribution. For example, for $\gamma = .50$, $z_\gamma = .6745$; for $\gamma = .68$, $z_\gamma = 1.0$; and for $\gamma = .95$, $z_\gamma = 1.96$. From Equation 2, it is immediate that

$$\Pr\left[\bar{X} - z_\gamma \hat{\sigma}_{e(\bar{X})} \le \pi \le \bar{X} + z_\gamma \hat{\sigma}_{e(\bar{X})}\right] \approx \gamma, \tag{3}$$

where $\hat{\sigma}_{e(\bar{X})} = \sqrt{\bar{X}(1-\bar{X})/k}$ is the estimated standard error for $\bar{X}$, and the subscript "$e$" represents the error of measurement.

Equation 3 gives an interval for $\pi$, $\left(\bar{X} - z_\gamma \hat{\sigma}_{e(\bar{X})}, \bar{X} + z_\gamma \hat{\sigma}_{e(\bar{X})}\right)$, which has two endpoints that are random variables each of which is dependent upon $X$; they will be

denoted as $X_L$ and $X_U$. So, it can be said that, prior to data collection, the probability that the random interval $(X_L, X_U)$ includes the unknown parameter $\pi$ is $\gamma$. Suppose we have collected data, then the two endpoints are known and the particular realized interval $(x_L, x_U)$ either does or does not cover $\pi$. However, if many such intervals were constructed over repeated applications of an interval estimation procedure, about $(100\gamma)\%$ of them would cover the parameter (Feldt & Brennan, 1989). The obtained interval $I(\pi) = (x_L, x_U)$ is called a $(100\gamma)\%$ confidence interval for $\pi$, and $\gamma$ is the confidence coefficient. In the statistics literature, the interval $\left(\overline{x} - z_\gamma \hat{\sigma}_{e(\overline{X})},\ \overline{x} + z_\gamma \hat{\sigma}_{e(\overline{X})}\right)$ is often called the Wald confidence interval for $\pi$ because it is derived from the Wald test for $\pi$.

The fact that $(\overline{X} - \pi)/\sqrt{\pi(1-\pi)/k}$ has a $N(0,1)$ limiting distribution implies that $X$ is distributed approximately as $N[k\pi, k\pi(1-\pi)]$ as $k$ goes to infinity (Hogg & Tanis, 1993). One can obtain a confidence interval for $\tau$, $I(\tau)$, by multiplying the two endpoints of $I(\pi)$ by $k$, since $\tau = \pi k$. As a result, $I(\tau)$ has the form $\left(x - z_\gamma \hat{\sigma}_{e(X)}, x + z_\gamma \hat{\sigma}_{e(X)}\right)$, which is referred to here as the Wald confidence interval for $\tau$, where $\hat{\sigma}_{e(X)} = \sqrt{k\overline{x}(1-\overline{x})} = \sqrt{x(k-x)/k}$ is the standard error for $X$. The confidence intervals expressed in two different metrics (i.e., total score vs. mean score) should not be confused.

***Conditional Confidence Intervals Using Conditional SEMs***

The rationale for conditional confidence intervals to be discussed here clearly parallels the rationale for the Wald confidence intervals discussed previously. The only difference is that conditional confidence intervals use an unbiased estimate of SEMs, which is called the Lord's SEM in the measurement literature. Under the binomial error model, Lord (1955, 1957) provided an estimated raw-score SEM for an examinee with $x$ items correct:

$$\hat{\sigma}_{e(X)_C} = \sqrt{\frac{x(k-x)}{k-1}} = \sqrt{\frac{k}{k-1}}\sqrt{\frac{x(k-x)}{k}}, \tag{4}$$

where $\sqrt{k/(k-1)}$ is a bias-correction factor to remove the bias in the variability of the sample. Then, the approximate confidence interval for the examinee has exactly the same form as the Wald confidence interval except that $\hat{\sigma}_{e(X)_C}$ is now used:

$$I_c(\tau) = \left(x - z_\gamma \hat{\sigma}_{e(X)_C}, x + z_\gamma \hat{\sigma}_{e(X)_C}\right). \tag{5}$$

It is well known that the normal approximation to the binomial distribution works best when $k$ is large and $\pi$ is close to .5, and many authors have suggested rules of thumb (see Leemis & Trivedi, 1996) for appropriate use of the normal approximation. For example, Hogg and Tanis (1993) considered $k$ sufficiently large if $k\pi \geq 5$ and $k(1-\pi) \geq 5$, or a $k$ of at least 30 in all cases. Note also that $\hat{\sigma}_{e(X)_C}$ would be zero for examinees with zero or perfect scores. Thus, it is very likely that the actual coverage probabilities for those examinees will be lower than the nominal coverage level.

***Traditional Confidence Intervals Using Overall SEM***

It has been customary to construct a $(100\gamma)\%$ confidence interval for $\tau$ using the normal approximation in conjunction with a strong assumption of the same SEM for all examinees. The traditional overall confidence interval is constructed as

$$I_o(\tau) = \left(x - z_\gamma \hat{\sigma}_{e(X)_O}, x + z_\gamma \hat{\sigma}_{e(X)_O}\right), \tag{6}$$

where $\hat{\sigma}_{e(X)_O}$ is the estimated overall SEM. The estimated overall raw-score SEM can be obtained by $\hat{\sigma}_{e(X)_O} = \sqrt{\sum \hat{\sigma}^2_{e(X)_C}/N}$, which is the square root of the average of Lord's error variances for all $N$ examinees in the sample (Brennan, 1996; Brennan & Kane, 1977). [In the terminology of generalizability theory, this is a $\Delta$-type error variance.]

*Score Confidence Intervals*

The score confidence interval was first discussed by Wilson (1927), and some studies have recommended it over other confidence intervals for $\pi$ (Ghosh, 1979; Agresti & Coull, 1998; Santner, 1998). The score confidence interval uses the population standard error of $\bar{X}$, rather than the estimator in Equations 2 and 3. The two endpoints of a score confidence interval are obtained from the fact that $(\bar{X}-\pi)/\sqrt{\pi(1-\pi)/k}$ has a $N(0,1)$ limiting distribution. Then, a probability statement similar to Equation 2 can be made:

$$\Pr\left[-z_\gamma \le \frac{\bar{X}-\pi}{\sqrt{\pi(1-\pi)/k}} \le z_\gamma\right] \approx \gamma. \tag{7}$$

Unlike Equations 2 and 3, Equation 7 is not directly solvable for $\pi$, but the solution is not very complicated. Equation 7 is equivalent to

$$\Pr\left[\frac{(\bar{X}-\pi)^2}{\pi(1-\pi)/k} \le z_\gamma^2\right] \approx \gamma. \tag{8}$$

The term in brackets in Equation 8 can be written as a quadratic equation for $\pi$. The two zeros of the quadratic equation in $\pi$ form the endpoints of the score confidence interval for $\pi$, $I_s(\pi)$. Now, the score confidence interval for $\tau$ is $k$ times the two endpoints of $I_s(\pi)$, which has the form

$$I_s(\tau) = k\left(\frac{x+z_\gamma^2/2}{k+z_\gamma^2} - z_\gamma\sqrt{\frac{x(k-x)/k+z_\gamma^2/4}{\left(k+z_\gamma^2\right)^2}}, \frac{x+z_\gamma^2/2}{k+z_\gamma^2} + z_\gamma\sqrt{\frac{x(k-x)/k+z_\gamma^2/4}{\left(k+z_\gamma^2\right)^2}}\right). \tag{9}$$

The midpoint of $I_s(\tau)$ can be rewritten as $x[k/(k+z_\gamma^2)]+k/2[z_\gamma^2/(k+z_\gamma^2)]$, which obviously falls between $x$ and $k/2$. This midpoint, in effect, shifts the midpoint of the

conditional and traditional confidence intervals, $x$, toward $k/2$. In addition, the multiplier of $z_\gamma$ in Equation 9 shows that the problem of zero standard errors for examinees with $x=0$ or $k$ is not present for this interval. The term under the square root will always result in a positive number regardless of the value of $x$.

***Bayes Confidence Intervals***

The normal approximation may not be very accurate when an examinee's observed proportion-correct score is near zero or one. An alternative estimate of $\pi$, rather than $\bar{x}$, would be a Bayes estimate $\tilde{\pi}=(x+\alpha)/(k+\alpha+\beta)$, which appears to give a more reasonable estimate than $\bar{x}$, especially for the extreme values of $X$ (Chen, 1990). The value of $\tilde{\pi}$ is the mean of the posterior distribution using the beta prior distribution with parameters $\alpha$ and $\beta$. Chen (1990) recommended $\alpha=\beta=z_\gamma^2/2$, which shrinks the individual's observed proportion-correct score to .5. The endpoints of this interval can be obtained by replacing $\bar{x}$ in the Wald confidence interval with $\tilde{\pi}$:

$$I_b(\tau)=k\left(\tilde{\pi}-z_\gamma\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{k}},\tilde{\pi}+z_\gamma\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{k}}\right). \tag{10}$$

Note that the midpoint of the Bayes interval with $\alpha=\beta=z_\gamma^2/2$ is the same as that of the score confidence interval, since $\tilde{\pi}=(x+\alpha)/(k+\alpha+\beta)=(x+z_\gamma^2/2)/(k+z_\gamma^2)$. As for the score confidence interval, the adjusted standard error for $\bar{x}$, $\sqrt{\tilde{\pi}(1-\tilde{\pi})/k}$, prevents the estimate from being zero for examinees with zero or perfect observed scores.

***Clopper-Pearson Exact Confidence Intervals***

The confidence intervals discussed so far are based on limiting distribution theory using $N(0,1)$. There are a few confidence intervals for the binomial parameter $\pi$, which use the binomial distribution (i.e., "exact" distribution for $X$) rather than the approximate normal distribution (Blyth & Still, 1983; Clopper & Pearson, 1934; Crow, 1956; Sterne,

1954). All exact confidence intervals are considered to be conservative in that they are guaranteed to have the coverage probability of *at least* $\gamma$ for any examinee with a particular true score (Santner, 1998). The conservatism of exact confidence intervals is due to the fact that the binomial distribution of $X$ is discrete, and thus an exact probability, say, .95, can not be attained (Agresti & Coull, 1998; Hogg & Craig, 1995).

Among several methods for constructing exact intervals, the Clopper-Pearson (1934) interval is the first and probably most widely known. The endpoints of the Clopper-Pearson confidence interval for $\tau$ are obtained as follows: $x_L$ is $k$ times the value of $\pi$ such that $\Pr(X \geq x \mid \pi, k) = (1-\gamma)/2$ and $x_U$ is $k$ times the value of $\pi$ such that $\Pr(X \leq x \mid \pi, k) = (1-\gamma)/2$, where

$$\Pr(X \geq x \mid \pi, k) = \sum_{j=x}^{k} \binom{k}{j} \pi^j (1-\pi)^{k-j} \tag{11}$$

and

$$\Pr(X \leq x \mid \pi, k) = \sum_{j=0}^{x} \binom{k}{j} \pi^j (1-\pi)^{k-j} . \tag{12}$$

The lower bound is taken to be 0 when $x = 0$, and the upper bound is taken to be 1 when $x = k$.

A simple method of solving Equations 11 and 12 involves using either the incomplete beta distribution or the $F$-distribution. Let $IB_\pi(\alpha, \beta)$ be the incomplete beta distribution for $\pi$ with parameters $\alpha$ and $\beta$. Then, $x_L$ is $k$ times the value of $\pi$ for which $IB_\pi(x, k-x+1) = (1-\gamma)/2$, and $x_U$ is $k$ times the value of $\pi$ for which $IB_\pi(x+1, k-x) = (1+\gamma)/2$. Using the $F$-distribution, the Clopper-Pearson interval is

$$I_e(\tau) = k\left(\{1+[\nu_1/\nu_2]F_{\nu_1,\nu_2,(1+\gamma)/2}\}^{-1}, \{1+[\nu_3/\nu_4]F_{\nu_3,\nu_4,(1-\gamma)/2}\}^{-1}\right), \tag{13}$$

where $\nu_1 = 2(k-x+1)$, $\nu_2 = 2x$, $\nu_3 = 2(k-x)$, and $\nu_4 = 2(x+1)$ are degrees of freedom, and $F_{(1\pm\gamma)/2}$ denotes the $(1\pm\gamma)/2$ th quantile of the $F$-distribution.

The Clopper-Pearson confidence interval is considered to be appropriate even for a small $k$. However, the actual coverage probability of this interval can be much larger than the nominal confidence level due to its conservatism.

***Credibility Intervals***

The statistical method of inference underlying credibility intervals is Bayesian statistics. The Bayesian estimation approach takes into account both the test score and any prior knowledge about the examinee's true score. Equation 1 enables us to compute the probabilities of various observed scores for a known value of $\pi$. Conversely, the Bayesian approach considers the problem of inferring the value of $\pi$ given $X = x$. In Bayesian statistics, all the information for making inferences about an examinee's true score is contained in the conditional distribution of the examinee's true (proportion-correct) scores given the observed score. Note that we now consider $\pi$ as a possible value of the random variable $\Pi$ rather than a constant for an examinee. Presumably, $\Pi$ is a continuous variable with an interval of $0 \le \Pi \le 1$.

The conditional distribution of $\Pi$ given $X = x$, $g(\pi \mid x)$, is called the posterior distribution. Let $f(x \mid \pi)$ denote the conditional probability density function of $X$, given $\Pi = \pi$. The goal of the Bayesian inference is to obtain the posterior distribution, $g(\pi \mid x)$, using a subjectively selected prior distribution, $h(\pi)$, and $f(x \mid \pi)$. A $(100\gamma)\%$ credibility interval is constructed by taking the $(1-\gamma)/2$ and $(1+\gamma)/2$ quantiles of the posterior distribution. An interval constructed in this way is sometimes referred to as a central or an equal-tailed credibility interval (Novick & Jackson, 1974).

According to Bayes theorem,

$$g(\pi \mid x) \propto f(x \mid \pi) h(\pi), \tag{14}$$

where the symbol $\propto$ is read "proportional to". The conditional observed score distribution, $f(x \mid \pi)$, is already known as the binomial model, and the beta distribution, $B(\alpha, \beta)$, is typically used as the prior density. The beta distribution has two parameters $\alpha$ and $\beta$, and by varying the two parameter values, one can obtain a family of beta densities whose functional form is very similar to that of the binomial distribution. Due to their similar density forms, the two distributions combine in a very convenient way. Using $B(\alpha, \beta)$ as the prior density, $h(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}$ and $f(x \mid \pi) \propto \pi^{x}(1-\pi)^{k-x}$. Thus,

$$g(\pi \mid x) \propto \pi^{x+\alpha-1}(1-\pi)^{k-x+\beta-1}. \tag{15}$$

It is obvious from Equation 15 that the posterior distribution has the form of another beta distribution with parameters $x+\alpha$ and $k-x+\beta$.

The endpoints of a $(100\gamma)\%$ credibility interval for $\pi$ are computed using the incomplete beta distribution with parameters $x+\alpha$ and $k-x+\beta$. That is, $x_L$ is the value of $\pi$ such that $IB_{\pi}(x+\alpha, k-x+\beta) = (1-\gamma)/2$, and $x_U$ is the value of $\pi$ such that $IB_{\pi}(x+\alpha, k-x+\beta) = (1+\gamma)/2$. A $(100\gamma)\%$ credibility interval for $\tau$ is obtained by multiplying $k$ by the two endpoints.

When the beta prior has parameters $\alpha = \beta = 1$, it is a uniform distribution on the interval [0,1] implying that all values of $\Pi$ are equally likely. This particular prior distribution is often called a non-informative prior. For the non-informative prior, the posterior distribution is $B(x+1, k-x+1)$. In this paper, the credibility interval for $\tau$ using the non-informative prior, $I_p(\tau)$, is compared with the confidence intervals previously described.

It is important to recognize that credibility intervals differ from confidence intervals in terms of logical interpretation. In a $(100\gamma)\%$ credibility interval, $(100\gamma)\%$ is the probability attached to the particular interval obtained for an examinee (Novick &

Jackson, 1974). As such, we can make a probability statement that the probability of an examinee's true score falling between the two endpoints of a credibility interval is $(100\gamma)\%$. It is the direct use of the distribution of $\Pi$ (i.e., posterior distribution) rather than the observed score distribution that makes possible the probability statement. By contrast, for a $(100\gamma)\%$ confidence interval, the probability attaches to the interval estimation method, not to the particular realized interval (Novick & Jackson, 1974). A confidence interval is constructed based on the observed score distribution, and the particular interval either does or does not contain the true score. As mentioned earlier, a confidence interval is typically interpreted as follows: if the interval estimation method were applied an infinitely large number of times, it would produce $(100\gamma)\%$ intervals that cover the true score.

## Intervals for Scale Scores

In most testing programs, raw scores typically are transformed to scale scores for the purposes of reporting and making decisions about examinees. Thus, if intervals are to be reported, they would be most informative if expressed in terms of scale scores. Two general procedures are considered in this paper for constructing intervals for scale scores. The first, the normal approximation method, might be used for scale scores in conjunction with conditional scale-score SEMs or overall scale-score SEMs. The second, the endpoints conversion method, provides scale-score counterparts of any raw-score interval by converting the lower and upper endpoints of an interval for $\tau$ to corresponding scale scores according to the functional relationship between raw scores and scale scores.

### *Normal Approximation*

Let $S$ be a random variable for scale scores that are transformed from raw scores, $X$, using the transformation function, $u(X)$. Let $\xi$ and $\sigma_{e(S)}$ denote the true scale score

and the scale-score SEM for an examinee, respectively. Here, $\xi$ is defined as a mean of scale scores obtained over repeated measurements, that is

$$\xi = \mathbf{E}[u(X)] = \sum_{i=0}^{k} u(i) \Pr(X = i \mid \pi), \tag{16}$$

where $\mathbf{E}$ is the expectation operator and $\Pr(X = i \mid \pi)$ is given in Equation 1.

Suppose $u(X)$ is a linear transformation function, $u(X) = A(X) + B$. Then, the shape of the conditional distributions for $X$ and $S$ will be the same, which makes it sensible to use the normal approximation for scale scores to the extent that it is sensible for raw scores. More specifically, if the limiting distribution of $X$ is $N[k\pi, k\pi(1-\pi)]$, the limiting distribution of $S$ is $N[Ak\pi + B, A^2 k\pi(1-\pi)]$. For a linear transformation, $\sigma_{e(S)}$ is simply $A\sigma_{e(X)}$. Using Lord's SEM (i.e., Equation 4) as the raw-score SEM, the conditional confidence interval for $\xi$ for an examinee with an observed scale score $s$ is

$$I_c(\xi) = \left(s - z_\gamma \hat{\sigma}_{e(S)_c}, s + z_\gamma \hat{\sigma}_{e(S)_c}\right), \tag{17}$$

where $\hat{\sigma}_{e(S)_c} = A\sqrt{x(k-x)/(k-1)}$. The overall confidence interval for $\xi$, under the linear transformation, can also be constructed as

$$I_o(\xi) = \left(s - z_\gamma \hat{\sigma}_{e(S)_o}, s + z_\gamma \hat{\sigma}_{e(S)_o}\right), \tag{18}$$

where $\hat{\sigma}_{e(S)_o} = \sqrt{\sum \hat{\sigma}^2_{e(S)_c} / N}$. The same value of $\hat{\sigma}_{e(S)_o}$ is used for all examinees.

Note that the normal approximation will result in exactly the same coverage probabilities for the linearly transformed scale scores and corresponding raw scores. For a linear transformation, $\xi = \mathbf{E}[u(X)] = \mathbf{E}[AX + B] = A\mathbf{E}(X) + B = u[\mathbf{E}(X)] = u(\tau)$, which clearly indicates that the transformation parameters $A$ and $B$ are for both observed and true scores. If and only if $x_L < \tau < x_U$, then $Ax_L + B < A\tau + B < Ax_U + B$, provided

that $A$ is a positive value. Consequently, the coverage probabilities for $I(\xi)$ and $I(\tau)$ will be the same under a linear transformation--the same argument applies to both conditional and overall confidence intervals. For non-linear transformations, however, the coverage probabilities for $I(\xi)$ and $I(\tau)$ will not be identical, because $\xi = \mathbf{E}[u(X)] \neq u[\mathbf{E}(X)]$, in general. In other words, the relationship between $X$ and $S$ is not the same as the relationship between $\tau$ and $\xi$ for a non-linear transformation.

A bigger concern about the normal approximation approach for the non-linearly transformed scale scores is that the assumption of the limiting normal distribution may not hold, because the non-linearity distorts the shape of the conditional distribution for $S$. Even though the normal approximation might work reasonably well for "moderate" non-linear transformations, it may not be appropriate for "severe" non-linear transformations. This paper considers four different types of non-linearly transformed scale scores, each of which has a different degree of non-linearity; and evaluates the performance of the normal approximation when applied to the various scale scores.

For a non-linear transformation, $\sigma_{e(S)}$ is not simply $A\sigma_{e(X)}$ because the slope parameter $A$ changes along the score scale. There exist several procedures for estimating conditional scale-score SEMs (CSSEMs) when $u(X)$ is non-linear. In this paper, a method called the binomial procedure (Brennan & Lee, 1997, 1999) is employed. The binomial procedure provides $\hat{\sigma}_{e(S)}$, which can be viewed as a scale-score analogue of Lord's SEM:

$$\hat{\sigma}_{e(S)_C} = \sqrt{\frac{k}{k-1}}\sqrt{\sum_{i=0}^{k}[u(i)]^2 \Pr(X=i \mid \pi=\bar{x}) - \left[\sum_{i=0}^{k} u(i)\Pr(X=i \mid \pi=\bar{x})\right]^2}. \qquad (19)$$

The conditional and overall confidence intervals for non-linearly transformed scale scores, respectively, can be constructed by Equations 17 and 18 using $\hat{\sigma}_{e(S)_C}$ in Equation 19.

***Endpoints Conversion***

As discussed in the previous section, one problem with the normal approximation method to constructing confidence intervals for non-linearly transformed scale scores arises due to the direct use of the conditional scale-score distributions for which the normality assumption seemingly does not hold. Another approach to constructing intervals for scale scores presented in this section is free from such a problem because it does not assume any distributional form of the scale scores. The method here called "endpoints conversion" finds the endpoints of scale-score intervals by converting the endpoints of raw-score intervals through a functional relationship between raw and scale scores. In effect, the scale-score counterpart of any raw-score interval can be obtained by the endpoints conversion method.

There seem to be at least two functional relationships that can be used for converting the endpoints. Obviously, the actual observed score transformation, $u(X)$, could be used, with which the endpoints for the scale-score counterpart of a raw-score interval are obtained as

$$I_u(\xi) = \left(u[x_L],\ u[x_U]\right), \tag{20}$$

where $x_L$ and $x_U$ are the two endpoints of the raw-score interval. However, there is a complexity. To use a preexisting conversion table, $u(X)$, it must be assumed that the transformation is a continuous function, because $x_L$ and $x_U$ are often non-integer values and the corresponding scale scores can not be read directly from the conversion table. Thus, an interpolation procedure is usually needed.

Another alternative is to use the relationship between true scores and true scale scores. Let $v$ denote the transformation function from $\tau$ to $\xi$ such that $\xi = v(\tau)$. Once the true-score conversion function, $v$, is determined, the two endpoints of a raw-score interval, $x_L$ and $x_U$, are substituted for $\tau$. For the case considered in this paper, $v$ is

defined by Equations 1 and 16, and the two endpoints of the scale-score counterpart can be obtained by substituting $x_L / k$ and $x_U / k$ (if $x_L$ and $x_U$ are in the total score metric) for $\pi$ in both equations. Let us express the resultant scale-score interval as:

$$I_v(\xi) = \left(v[x_L],\ v[x_U]\right). \tag{21}$$

Note that the notation of $I_u(\xi)$ and $I_v(\xi)$ indicates that they are generic intervals. That is, the endpoints of the intervals can be obtained from any raw-score endpoints, and different raw-score endpoints will result in different endpoints for $I_u(\xi)$ and $I_v(\xi)$. When the true-score conversion, $v$, is used, the coverage probability of the scale-score counterpart, $I_v(\xi)$, will be equal to that of the raw-score counterpart regardless of whether the raw-to-scale score transformation is linear or non-linear, because the endpoints and true scores are converted through the same conversion function, $v$. By contrast, the coverage probability of $I_u(\xi)$ will not be the same as that of the raw-score counterpart, because the endpoints and true scores are converted through different conversion functions, $u$ and $v$. In general, the true score conversion approach seems more reasonable. It is consistent with the fact that the endpoints are in the metric of true scores. The endpoints are almost always non-integer values and are compared with the true score to determine the coverage. Moreover, the same coverage probability for raw- and scale-score intervals seems appealing in practice.

The two approaches will produce very similar results, however, for a nearly one-to-one transformation. Figure 1 depicts the two conversions for the four types of scale scores for ITBS Vocabulary ($k = 34$), Form K, Level 10. The observed-score conversions are the ones that are used operationally for the test, and the true-score conversions were computed using Equation 16. Note that Equation 16 will result in zero values for $\xi$ when $\tau$ is either zero or $k$. Hence, the maximum and minimum values of $\xi$ were set to equal the maximum and minimum values of the scale scores in the observed-

score conversions. The label "Raw Score" for the horizontal axis in Figure 1 should be interpreted as either the true score or observed score depending upon what type of conversion is under consideration. Notice that the two conversions are extremely similar for DSSs, GEs, and PRs, largely because they are one-to-one functions throughout most of the score ranges. The largest difference is found in the raw-to-ST conversion, where many observed raw score points are converted to the same stanine point. The true-score conversions are strictly increasing functions (i.e., one-to-one at any score point), and appear to be smoother than the observed-score conversions, in general. This paper considers the true-score conversions only. Some results based on the observed-score conversions are discussed by Lee (1998).

## Numerical Example

The interval estimation procedures discussed in the previous sections are illustrated using the same test with the conversion table shown in Figure 1. Note that the confidence intervals using the overall SEMs are not considered in this example because they require actual examinee data. Table 1 displays actual endpoints of the nominal 68% raw and DSS intervals at five different score points: $x = 5$ $(DSS = 141)$; $x = 10$ $(DSS = 165)$; $x = 17$ $(DSS = 187)$; $x = 25$ $(DSS = 207)$; and $x = 30$ $(DSS = 229)$. The shaded areas in Table 1 indicate the endpoints of DSS intervals obtained through the endpoints conversion method with the true-score conversions, $I_v(\xi)$. The lower panel of Table 1 shows the actual raw-to-DSS conversion for the test, which is used to calculate the DSS intervals (Hoover, Hieronymus, Frisbie, & Dunbar, 1993b). Readers can verify the results reported in Table 1 using this conversion table.

The endpoints of $I_s(\tau)$ and $I_b(\tau)$ tend to be closer to each other than to any other intervals for the nominal level of 68%. Moreover, it can be verified that the midpoints of $I_s(\tau)$ and $I_b(\tau)$ are always the same and shifted toward $k/2$ from the midpoint of $I_c(\tau)$, which equals $x$. That is, the midpoint of $I_s(\tau)$ and $I_b(\tau)$ is larger than $x$ when

$x < k/2$ and smaller than $x$ when $x > k/2$. The midpoint shift toward $k/2$ is also observed for $I_e(\tau)$ and $I_p(\tau)$. At $x = k/2$, which is 17, the midpoints of all five raw-score intervals are equal to 17. Note that all procedures produce different endpoints, although rounding may cause some endpoints to appear equal.

For the DSS intervals, only $I_c(\xi)$ has midpoints that are equal to the observed DSS scores. The midpoint of $I_v(\xi)$ converted from $I_c(\tau)$ does not necessarily equal the observed DSS score because of the non-linearity of the raw-to-DSS transformation. Likewise, the midpoints of the DSS counterparts of $I_s(\tau)$ and $I_b(\tau)$ are not necessarily the same although their raw-score counterparts have the same midpoints. Another important property of the various interval estimation methods is the lengths of the intervals. The lengths of each interval across the score scales are plotted in Figure 2.

First notice in Figure 2 that the patterns of the various interval lengths are very similar. Actually, the shapes of the interval lengths reflect the shapes of the conditional SEMs (presented later in Figure 8)--large (small) conditional SEMs lead to wide (narrow) intervals. The irregular pattern of the DSS interval lengths is due to the non-linear character of the raw-to-DSS transformation. Notice also that the lengths of $I_e(\tau)$ for both the raw and DSS scores are remarkably larger than the other intervals throughout the score range, which, as discussed later, is closely related to the fact that the coverage probabilities of $I_e(\tau)$ are exceptionally large. The lengths of the intervals except for $I_e(\tau)$ do not appear to be very different except at both ends of the score scales. Especially, $I_c(\xi)$ and $I_v(\xi)$ converted from $I_c(\tau)$ exhibit very similar lengths of the DSS intervals even though the endpoints of the two intervals are not very close to each other (see Table 1). Note that the lengths of $I_v(\xi)$ converted from $I_c(\tau)$ and $I_c(\xi)$ are zero for the zero and perfect raw and corresponding DSS scores, which is caused by the zero estimated conditional SEMs. The approximate ascending order of the raw-score interval lengths, mainly in the middle of the score scale, is $I_p(\tau)$, $I_s(\tau)$, $I_b(\tau)$, $I_c(\tau)$, and $I_e(\tau)$. The same ordering applies to the corresponding DSS intervals. All other

things being equal (such as the same coverage probability), a method with narrow intervals would be preferred.

## Simulation Study

Since all procedures discussed in this paper are associated with the binomial distribution of errors in one way or another, a simulation was conducted based on a model called the beta-binomial model (Keats & Lord, 1962; Lord & Novick, 1968), which assumes that errors are distributed binomially. The beta-binomial model is known to fit many observed score distributions very well. In order to generate random data that are as realistic as possible, a real test data set initially was used for specifying the simulation conditions. This simulation study used data from Level 10, Form K of the Vocabulary subtest ($k = 34$) in ITBS--a random sample of 3000 examinees at grade 4 (Level 10) was selected from the 1992 Spring standardization sample.

Under the beta-binomial model, the conditional distribution of $X$ given $\pi$ is binomial, and $\pi$ is distributed as beta with parameters $\alpha$ and $\beta$, $B(\alpha,\beta)$. Let $\hat{\rho}_{21} = [k/(k-1)][1-\hat{\mu}(k-\hat{\mu})/kS^2]$ denote KR21 reliability, where $\hat{\mu}$ and $S$ are the mean and standard deviation of the test scores for the 3000 examinees. The parameters $\alpha$ and $\beta$ were estimated using the following formulas (see Huynh, 1976; Jarjoura, 1985):

$$\hat{\alpha} = \hat{\mu}\left(\frac{1}{\hat{\rho}_{21}} - 1\right)$$

and

$$\hat{\beta} = (k-\hat{\mu})(\frac{1}{\hat{\rho}_{21}} - 1). \qquad (22)$$

Equation 22 yielded $\hat{\alpha} = 3.4$ and $\hat{\beta} = 1.9$, which suggests that the distribution of $\pi$ is a bit negatively skewed. Negatively-skewed distributions of test scores are typical

for many standardized achievement tests. Treating the two parameter estimates as population parameters, true proportion-correct scores, $\pi$, were generated for 1000 simulees from $B(3.4,1.9)$. In generating random beta deviates, the acceptance-rejection method was employed as described in Mooney (1997, pp. 25-30). For each simulee, the conditional observed score distribution, $\Pr(X = x \mid \pi)$, was computed using Equation 1, and the true number-correct score and true scale score were computed as $\tau = k\pi$ and $\xi = \sum_{i=0}^{k} u(i) \Pr(X = i \mid \pi)$. Then, the following steps were executed:

1. A set of $k = 34$ random 0/1 item responses for each of the 1000 simulees was generated by comparing a uniform random deviate $r$ to $\pi$ for 34 times. If $r \leq \pi$ then a score of one was assigned to the item, otherwise a score of zero was assigned.
2. All interval estimation procedures were applied to the simulated data, and intervals were constructed for each of the 1000 simulees.
3. For each simulee, it was determined whether each interval contained the simulee's true (scale) score. If an interval covered the parameter, $C_r = 1$, otherwise, $C_r = 0$.
4. The above steps were replicated $R = 1000$ times and $\sum_{r=1} C_r$ was calculated for each simulee, which represents the empirical number of times that the intervals obtained from repeated measurements include the true (scale) score. The actual coverage probability was computed for each simulee, each interval estimation procedure, and each type of scale score.

The simulation procedure was replicated for three different nominal confidence levels: 95%, 68%, and 50%. The actual coverage probabilities obtained through the simulation were compared to these three nominal levels. These three nominal levels were used in the previous study by Jarjoura (1985).

Finally, note that the number of items in the original test was 34. To examine the effect of the number of items in the test, the whole simulation was repeated for $k = 17$. In all other respects, the population characteristics were exactly the same. A shorter version of the conversion table was created somewhat arbitrarily by removing the even-numbered rows in the original conversion table. Consequently, the patterns of the transformations for the shorter test were remarkably similar to those of the original test. Figures 3 and 4 show the plots of the transformations for the two tests.

## Results

### *Nominal 95% Intervals*

Table 2 contains averages and standard deviations of actual coverage probabilities for the nominal 95% intervals. The averages and standard deviations were computed based on 1000 simulees' actual coverage probabilities, and the averages are, in fact, the actual coverage probabilities over 1,000,000 intervals (1000 simulees times 1000 replications).

For the raw-score intervals, the score confidence interval, $I_s(\tau)$, and the credibility interval with the non-informative prior, $I_p(\tau)$, appear to show the actual coverage probabilities close to the nominal level of .95 with relatively small standard deviations. The Bayes confidence interval, $I_b(\tau)$, and the Clopper-Pearson exact confidence interval, $I_e(\tau)$, tend to be somewhat conservative (i.e., larger actual coverage probabilities than the nominal level). Recall that $I_e(\tau)$ is supposed to have coverage probabilities that are constantly bounded below by the nominal confidence level. The somewhat conservative coverage probabilities of $I_b(\tau)$ are consistent with the previous results reported by Agresti and Coull (1998). The conditional confidence interval using Lord's SEM, $I_c(\tau)$, yielded the actual coverage probability that is too small and has the largest standard deviation, which is likely due to zero estimated SEMs for $x = 0$ or $k$.

With a zero SEM, the width of $I_c(\tau)$ is zero and the actual coverage probability can be too low. Apparently, the overall confidence interval using overall SEMs, $I_o(\tau)$, seems to perform better than $I_c(\tau)$. It is not necessarily true, however, that a procedure showing a better overall coverage probability is more accurate across all levels of the score scale. Some procedures might be more accurate than other procedures near the middle of the score scale but less accurate at extremes. More discussion about this issue is presented later.

The shaded areas in Table 2 and all the subsequent tables represent the coverage probabilities for the scale-score counterparts of each of the six raw-score intervals, $I_v(\xi)$, obtained by the endpoints conversion method with the true-score conversions. Note that the coverage probabilities for $I_v(\xi)$ are exactly the same as those for corresponding raw-score intervals regardless of the types of scale scores. The last two columns of Table 2 are for the conditional and overall scale-score confidence intervals, $I_c(\xi)$ and $I_o(\xi)$. Clearly, $I_o(\xi)$ provides better actual coverage probabilities and smaller standard deviations than $I_c(\xi)$. As for the conditional raw-score confidence intervals, zero estimated CSSEMs at both ends of the score scales are a major problem with $I_c(\xi)$. The results suggest that $I_v(\xi)$ associated with "good" raw-score intervals such as the score and Bayes confidence intervals would be preferable to $I_c(\xi)$ and $I_o(\xi)$ for the nominal 95% intervals.

The plots of actual coverage probabilities for raw-score intervals are shown in Figure 5. Each dot represents the coverage probability for a single examinee. Notice that the actual coverage probability varies across the levels of the score scale. The coverage probabilities of $I_o(\tau)$ display a U-shape trend indicating that the actual coverage probability for an examinee would be either too large or too small depending upon where the examinee's true score is located on the continuum, except for the regions where the reference line crosses the function of the actual coverage probabilities. As discussed later (i.e., Figure 8), this is consistent with the fact that the pattern of the conditional raw-score

SEMs is an inverted U-shape, and thus the average SEM would be too large near both extremes and too small in the middle of the score scale. $I_o(\tau)$ may not be adequate for reporting individual-level confidence intervals in practice.

By contrast, the coverage probabilities for $I_c(\tau)$ are close to .95 in the middle of the score range, and tend to decline with increases in the absolute deviation of the raw score from the mid-score point. These results are consistent with the conventionally known fact that the normal approximation for the binomial parameter works best for $\pi$ values around .5, which is, in the present case, equivalent to the true score of 17. As discussed earlier, $I_c(\tau)$ shows a large drop in the coverage probabilities at the right end of the score scale approaching zero. The similar drop would have been noticed at extremely low true scores if the simulated data had contained enough data points at the region. Compared to $I_o(\tau)$, the coverage probabilities of $I_c(\tau)$, however, are fairly consistently closer to the nominal level throughout most of the score scale.

The actual coverage probabilities of $I_s(\tau)$ and $I_p(\tau)$ tend to be reasonably well scattered around the reference line. Notice, however, that both $I_s(\tau)$ and $I_p(\tau)$ show a little drop at the right extreme. For $I_p(\tau)$, the endpoints of a 95% interval when $x=33$ and $x=34$ (i.e., a perfect score) are (28.928, 33.762) and (30.599, 33.976), respectively. Suppose an examinee has a true score of 33.8. Whenever the examinee's observed score is less than perfect, the interval will not cover the examinee's true score. The upper endpoint of $I_p(\tau)$ when $x=k-1$ is the lower bound of a range of $\tau$ values that falls in the interval only when $x=k$. The simulated data actually had three simulees with the true scores greater than 33.762, which exactly matches the number of dots in the plot that are far below the nominal level at the right end of the score scale. Although the simulated data do not contain such cases, a similar remark can be made for $\tau$ values near zero. There is also a range of $\tau$ values that can be covered in the interval only when $x=0$. The range is bounded above by the lower endpoint of the interval when $x=1$, which is .238 for $k=34$. The credibility interval has an additional problem. Note that

the upper endpoint of $I_p(\tau)$ is 33.976 when $x=34$. Thus, the coverage probability for an examinee with a true score greater than 33.976 will be zero necessarily. The present data do not have such high true score values. As discussed later, however, this actually happens with lower nominal confidence levels when the width of the interval gets narrower. The range of $\tau$ values for which the actual coverage probability is necessarily zero gets smaller as the number of items increases.

The score confidence interval has a similar, but less serious problem. The endpoints of the 95% score confidence interval when $x=33$ is (28.929, 33.823). Note that the upper endpoint is larger than the corresponding value for $I_p(\tau)$. Again, the upper endpoint 33.823 is the lower bound of the range of $\tau$ values that falls in the interval only when $x=34$. There is only one simulee as shown in the plot who has a true score greater than 33.823. However, $I_s(\tau)$ does not have the problem of zero coverage probability for extremely high true scores as does $I_p(\tau)$, because the upper endpoint of $I_s(\tau)$ when $x=k$ is always $k$.

The actual coverage probabilities of $I_b(\tau)$ and $I_e(\tau)$ are almost uniformly larger than the nominal level along the entire score scale, with $I_e(\tau)$ being somewhat more conservative. Compared to $I_p(\tau)$ and $I_s(\tau)$, the range of $\tau$ values that is covered by the exact confidence interval only when $x=k$ is very small. This range does not even exist for $I_b(\tau)$. When $x=33$, the upper endpoint of $I_e(\tau)$ is 33.975, but the upper endpoint of $I_b(\tau)$ is 34.426, which is greater than the maximum true score, $k$. The upper endpoint of $I_e(\tau)$ when $x=k$ is set equal to $k$. The upper endpoint of $I_b(\tau)$ when $x=k$ is allowed to be greater than $k$--the limit approaches $k$ from above as $k$ goes to infinity.

The actual coverage probabilities for $I_c(\xi)$ and $I_o(\xi)$ are plotted in Figures 6 and 7. Notice that, as for the raw score case, the coverage probabilities for the conditional scale-score confidence intervals are very low near the right extreme because of zero estimated CSSEMs when $x=k$. Unlike the results for $I_c(\tau)$ and $I_o(\tau)$, however, the patterns of the coverage probabilities for the scale-score confidence intervals across the

score scales tend to be irregular largely because of the non-linearity in the transformations. There are at least two potential sources of inaccuracy due to the non-linearity, which causes the coverage probabilities for the scale-score confidence intervals based on the normal approximation to deviate from the nominal levels: (a) bias in the estimated CSSEMs and (b) violation of the normality assumption.

The degree of bias in the estimated CSSEMs can be evaluated by comparing them with the true CSSEMs. Since the parameter $\pi$ is known for each simulee, the true SEMs can be computed. The true raw-score SEM for an examinee under the binomial error model is $\sigma_{e(X)} = \sqrt{k\pi(1-\pi)}$, and the true CSSEM is

$$\sigma_{e(S)} = \sqrt{\sum_{i=0}^{k}[u(i)]^2 \Pr(X=i \mid \pi) - \left[\sum_{i=0}^{k} u(i) \Pr(X=i \mid \pi)\right]^2}. \qquad (23)$$

Figure 8 displays the true and mean estimated SEMs (over replications) for the raw and scale scores. The shape of the conditional raw-score SEMs is a concave-down parabola (Brennan, 1996, 1998; Feldt et al., 1985), and there does not seem to exist any noticeable bias in the estimated SEMs. The mean estimated overall raw-score SEM is a constant and an overestimate for examinees with very low and high true scores, but an underestimate for examinees in the middle of the true score distribution.

The CSSEMs typically are irregular depending upon the pattern of non-linear transformations (Brennan & Lee, 1997, 1999; Feldt & Qualls, 1998; Kolen, Hanson, & Brennan, 1992). The CSSEMs are larger, in general, at the score points where the slope is steeper. However, in many cases, the estimated CSSEMs tend to be small at both extremes of the score scales regardless of the degree of the slope because the conditional raw-score SEMs are too small at the extremes (Brennan & Lee, 1997, 1999). Lee, Brennan, and Kolen (1998, 2000) reported that the estimated CSSEMs tended to be biased, and the direction of bias was related to the magnitude of the CSSEMs along the

score scale. As seen in Figure 8, the CSSEMs tend to be overestimated near the middle values of the true scale scores, at which the CSSEMs are the local minima. However, the CSSEMs are underestimated near the true scale scores showing the local maximum values of the CSSEMs. It may be noticed in Figures 6 and 7 with conditional DSS and GE intervals that the most accurate coverage probabilities are associated with the scale score values at which the pattern or slope of the transformations change (i.e., inflection points) shown in Figure 1. Indeed, Lee et al. (1998, 2000) found that the degree of bias in the estimated CSSEMs is smallest near the inflection points. Also, notice that the actual coverage probabilities for the PR intervals tend to be less irregular than the other scale score results, because the raw-to-PR transformation is nearly linear throughout most of the score scale. The constant estimated mean overall scale-score SEMs shown in Figure 8 produces bias throughout the score scales. Figure 9 shows the bias plots for the estimated SEMs.

Figure 10 provides plots of the nominal 95% confidence intervals based on the normal approximation using the true conditional SEMs computed by Equation 23. The actual coverage probabilities are nearly uniformly distributed around the reference line of .95. Comparing Figure 10 and Figures 5, 6, and 7 along with Figure 9 provides a general idea about the effect of bias in the estimated SEMs on the actual coverage probabilities. It seems evident that the patterns of the coverage probabilities depicted in Figures 5 (conditional and overall confidence intervals only), 6, and 7 tend to mirror, in general, the patterns of the bias functions shown in Figure 9. However, the actual coverage probabilities for the confidence intervals with conditional SEMs near very high and low true scores tend to be too small even though the estimated conditional SEMs do not exhibit any noticeably large bias. This appears to be caused by an SEM of zero when an observed raw score is equal to $k$.

How can the actual coverage probability be too small when there is no bias in the mean estimated SEM? Let us take an example. Consider a simulee with a true score of

33.831 and the true SEM of 0.410. For this simulee, a large number of confidence intervals will not contain the true score when $x = k$ because the estimated SEM is zero. However, the mean estimated SEM is close to the true SEM, because, for this simulee, $x$ is equal to 34 or 33 in most cases and the estimated SEM is either zero or 1.0 (see Equation 4), and thus, the average becomes close to the true value of 0.410.

The variability of the coverage probabilities in Figure 10 is indicative of the effect of the violation of the normality assumption. The degree of the violation seems to vary depending on the types of scale scores and the location of the true (scale) scores. Obviously, the results for ST show the largest variability. This issue is discussed in a greater detail in the next section of nominal 68% intervals.

***Nominal 68% Intervals***

The results for the nominal 68% intervals are summarized in Table 3. In general, the conditional confidence intervals for both raw and scale scores, $I_c(\tau)$ and $I_c(\xi)$, outperform the overall confidence intervals, $I_o(\tau)$ and $I_o(\xi)$, with respect to both the average and standard deviation of the actual coverage probabilities. Especially, $I_c(\tau)$ performed nearly as well as $I_p(\tau)$ and $I_b(\tau)$. For scale scores, $I_v(\xi)$ appears to provide better coverage probabilities than $I_c(\xi)$ and $I_o(\xi)$, when the endpoints of $I_v(\xi)$ are obtained from converting the endpoints of any raw-score intervals except for $I_o(\tau)$ and $I_e(\tau)$. Note that the coverage probabilities for $I_o(\xi)$ are larger than .68 for all four types of scale scores. The results also show that $I_s(\tau)$ works better than the others, which in turn, leads to the better performance of the scale-score counterparts of $I_s(\tau)$. The actual coverage probabilities of $I_e(\tau)$ are excessively large. In comparison with the results for the nominal level of .95, the standard deviations of the actual coverage probabilities for the 68% confidence intervals tend to be much larger. One reason might be that the coverage probabilities of 68% confidence intervals have more room to move up and down.

The results for 68% raw-score intervals are plotted in Figure 11. Clearly, $I_c(\tau)$ provides the coverage probabilities that are closer to the nominal level than $I_o(\tau)$ throughout most of the score scale. Also notice that the patterns of the actual coverage probabilities for $I_s(\tau)$, $I_b(\tau)$, and $I_p(\tau)$ are remarkably similar, except that $I_p(\tau)$ shows a zero coverage probability at the true score of near 34. The endpoints of a 68% credibility interval when $X = 34$ are (32.266, 33.831). Since $x = 34$ is the maximum number of items correct, the upper endpoint of the interval can not be greater than 33.831. Thus, the coverage probability of the credibility interval for an examinee with the true score greater than 33.831 will be zero regardless of the examinee's observed score, and the simulated data contain one simulee with such a high true score. A similar remark can be made for the other end of the score range.

Figures 12 and 13 depict the actual coverage probabilities for 68% scale-score confidence intervals. As noted in Table 3, $I_c(\xi)$ shows much better patterns for the actual coverage probabilities than $I_o(\xi)$, except for the ST results. The excessive variation in the coverage probabilities for both conditional and overall ST confidence intervals makes them totally unacceptable. Since the plot for the ST counterpart of, for example, $I_s(\tau)$ will be the same as the plot for $I_s(\tau)$ (i.e., the coverage probabilities for $I_v(\xi)$ are identical to those for the corresponding raw-score intervals), the endpoints conversion method clearly provides better coverage probabilities for STs.

Figure 14 contains plots for the nominal 68% confidence intervals based on the normal approximation using the true conditional SEMs. Since the true SEMs are used here, the variations in the actual coverage probabilities are mainly due to the violation of the normality assumption. It appears that the normality assumption does not hold very well for STs and PRs at both ends of the score scale. For the sake of argument, let us consider the PR case, and presume that the normality assumption holds fairly well across the entire range of the raw-score scale. As shown in Figure 3, the slope of the raw-to-PR transformation is almost linear along the score scale except for both extremes where the

transformation becomes flat. Thus, the PR distribution at very high or low score points will be much narrower than the raw-score distribution, which, in turn, will result in large actual coverage probabilities as seen in Figure 14.

*Nominal 50% Intervals*

The actual coverage probabilities for the nominal 50% intervals are presented in Table 4. The inferences that can be made from Table 4 are pretty much the same as those that can be made from Table 3 for the 68% intervals. Some minor differences include that the results for 50% intervals show slightly larger standard deviations than those of the 68% intervals, in general. Also, the better performance of the conditional confidence intervals than the overall confidence intervals becomes more salient. In addition, the coverage probabilities for $I_e(\tau)$ now tend to exceed the nominal level to a prohibitive degree.

Plots are provided in Figures 15 through 18. The coverage probabilities for $I_s(\tau)$ and $I_b(\tau)$ tend to get more similar as the nominal level decreases shown in Figure 15. The coverage probabilities for the nominal 50% scale-score intervals (Figures 16 and 17) are generally more variable than those for the higher nominal levels. In particular, the ST confidence intervals display overly variable coverage probabilities. The large variation in the coverage probabilities of the ST confidence intervals is primarily due to the many-to-one conversion characteristics of the raw-to-ST transformation (see Figure 3). Likewise, the coverage probabilities for PR intervals tend to be more variable than those for DSSs and GEs at both tails of the score scale. As shown in Figure 3, several raw-score points are converted to the same PR point at both ends of the raw-to-PR transformation. In addition, the coverage probabilities for the PR intervals are fairly flat, like those for the conditional raw-score intervals, which is associated with the approximately linear pattern of the raw-to-PR transformation throughout a wide range of the score scale. The actual coverage probabilities for the conditional 50% confidence intervals using the true

conditional SEMs (i.e., Figure 18) show patterns that are similar to those for the 68% intervals, except that the 50% ST intervals and PR intervals at both ends tend to produce very low coverage probabilities due to narrow intervals coupled with narrow scale score distributions.

***Intervals for the Half-Length Test***

The whole simulation study was repeated for a test with a smaller number of items ($k = 17$), and the results for the three nominal confidence levels are summarized in Tables 5, 6, and 7. A more meaningful interpretation of these results might be made through a comparison with the results for the original test. One apparent difference is that, with the shorter test, the standard deviations of the actual coverage probabilities for all interval estimation procedures are somewhat larger than those with the full-length test. As expected, the conditional confidence intervals performed worse with the shorter test--they produced coverage probabilities that are noticeably lower or higher than the nominal levels. Also notice that the overall coverage probabilities for $I_e(\tau)$ for the shorter test are larger than those for the longer test, which were themselves larger than the nominal levels. In general, the other three raw-score intervals tend to work slightly worse with the shorter test.

A series of figures is provided for the shorter test results: Figures 19 - 21 for 95%; Figures 22 - 24 for 68%; and Figures 25 - 27 for 50%. A few comments will suffice. The general patterns of the actual coverage probabilities are very similar to those for the full-length test. For the scale-score confidence intervals, the similarity might be due to the similar pattern of the transformations for the two tests as shown in Figures 3 and 4. The plots for the shorter tests show more white spaces between chunks of dots, however, which is due to discreteness. There are only 17+1 possible intervals, relative to the longer test for which there are 35 possible intervals.

## Discussion

Agresti and Coull (1998) recommended score confidence intervals for nearly any sample size and parameter value, and the results of the present study support this recommendation. One minor drawback of the score confidence interval is that, as discussed earlier, it would have an actual coverage probability that is far below the nominal confidence level for an examinee with the true proportion-correct score of near zero or one. As the number of items increases, however, the problem diminishes. On average, the score confidence intervals provided the actual coverage probabilities closest to the nominal levels regardless of the test length.

One interesting observation is that the actual coverage probabilities for the score confidence intervals in Figures 5, 11, and 15 are almost identical to those for the conditional raw-score confidence intervals using the true SEMs (Figures 10, 14, and 18). This appears to be related to the fact that the true SEM is defined in this paper based on the binomial model, and that the derivation of the endpoints for a score confidence interval involves use of the true SEMs (see Equations 7 and 8).

In general, credibility intervals with a non-informative prior worked very well. One conceptual advantage of the credibility intervals is that we can make a probabilistic statement about a particular interval. There seem to be two technical problems with the credibility intervals, however. One, as with the score confidence intervals, there are true score regions at which the actual coverage probabilities could be too low, and the regions tend to be slightly larger than those associated with the score confidence intervals. This problem diminishes as the test gets longer. Two, especially for lower nominal levels, the actual coverage probabilities can drop to zero for extremely high or low true scores. A practical solution to the second problem might be to set the upper endpoint of the interval equal to the maximum number-correct score for an examinee with a perfect score, and the lower endpoint equal to zero for an examinee with zero score.

The results of the present study clearly showed that the Clopper-Pearson exact confidence intervals give actual coverage probabilities that substantially exceeded the nominal confidence levels. The exact confidence intervals are useful, however, as a conservative procedure. That is, with these intervals we can be sure that intervals will, on average, have at least the desired coverage probability regardless of score levels. Of course, if it is desired to have coverage probabilities as close as possible to the specified level at all score points, then the score and credibility intervals would be preferable.

The performance of the Bayes confidence intervals was acceptable, and it worked especially well with the nominal levels of .68 and .50. One advantage of the Bayes confidence intervals is that it does not have the problem of seriously low coverage probabilities. Also, the form of the Bayes intervals is identical to the familiar confidence interval form of $x \pm (z_\gamma)SEM$ using the Bayes estimate in place of the mean observed score.

Users might still insist on using intervals that involve adding and subtracting estimated SEMs multiplied by a $z$-score, since they are very popular and easy to implement. In such cases, it is recommended that the conditional SEMs be used rather than the overall SEM, especially for moderate and small confidence levels. Though the traditional overall confidence intervals, on some occasions, provide the overall actual coverage probabilities closer to the nominal confidence level, the intervals using conditional SEMs tend to produce the actual coverage probabilities that are constantly closer to the nominal level across the wide range of the score scale. This recommendation is applicable to both raw- and scale-score confidence intervals.

In addition, note that the computation of the conditional confidence intervals is based on test data for a single examinee only, whereas the traditional confidence intervals using the overall SEM make use of data from other examinees. Therefore, the conditional confidence interval might be more appropriate for a uniquely identifiable examinee. For instance, a counselor dealing with an individual student (especially one

who is particularly challenged or able) may be well advised to use an interval such as $I_c(\tau)$, because it is based on the student's test data only. By contrast, a test publisher, not knowing individual examinees, may opt for reporting intervals such as $I_o(\tau)$ in test manuals.

The accuracy of the actual coverage probabilities for the conditional scale-score confidence intervals appears to depend upon (1) the pattern of transformation (i.e., slope), (2) the accuracy of the estimated CSSEMs, and (3) the transformation type (one-to-one or many-to-one). The pattern of the transformation is closely related to the accuracy of the estimated CSSEMs. Both the pattern and type of the transformation are important factors since they can distort the shape of the conditional scale score distributions and thus the normality assumption may not hold any more for the scale scores. Given the fact that the normal approximation works fairly well for the raw scores, the more severe the transformation, the more likely the normality assumption is violated for the scale scores.

When constructing intervals for scale scores, the results presented here suggest that the endpoints conversion method using the true-score conversion is preferable to the normal approximation approach. It is recommended that the normal approximation be used for scale-score confidence intervals only when the transformation is approximately linear. One comment on the true-score conversion should be made. In order to get the $\nu$ transformation, which converts true scores into true scale scores, we begin with obtaining the observed score distribution given $\tau$. Doing so requires a model, and in the present case, the binomial error model was used. Although we can use any psychometric model for the true-score conversion that is assumed to hold for our data, such as one based on item response theory, the actual coverage probability for the scale-score counterparts will always be the same as the raw-score coverage probability.

Although this paper employed the true-score conversion for the endpoints conversion method, the observed-score conversion (i.e., Equation 20) could be another alternative. For a nearly one-to-one transformation, the two conversion functions will be

very similar, and the resultant scale-score endpoints will be very similar as well. However, when the transformation is many-to-one, such as the raw-to-ST transformation considered in this paper, the true-score conversion would be smoother than the observed-score conversion, and provide somewhat better coverage probabilities (see Lee, 1998 for the results of the ST confidence intervals using the observed-score conversion). In general, the true score conversion approach would be preferred because it is consistent with the fact that the endpoints are on the metric of true scores, and it always produces the same coverage probability for the raw- and scale-score intervals. Depending upon the type of the transformation, the observed-score conversion approach might be preferred because it is relatively easy to implement using a simple interpolation procedure.

## References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*, 119-126.

Blyth, C. R., & Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association, 78*, 108-116.

Brennan, R. L. (1996). *Conditional standard errors of measurement in generalizability theory* (Iowa Testing Programs Occasional Paper No. 40). Iowa City, IA: University of Iowa.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22*, 307-331.

Brennan, R. L., & Kane, M. T. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika, 42*, 609-625.

Brennan, R. L., & Lee, W. (1997). *Conditional standard errors of measurement for scale scores using binomial and compound binomial assumptions* (Iowa Testing Programs Occasional Paper No. 41). Iowa City, IA: University of Iowa.

Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 59*, 5-24.

Chen, H. (1990). The accuracy of approximate intervals for a binomial parameter. *Journal of the American Statistical Association, 85*, 514-518.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*, 404-413.

Crow, E. L. (1956). Confidence intervals for a proportion. *Biometrika, 43*, 423-435.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*, 883-891.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education and Macmillan.

Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement, 33*, 141-156.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education, 11*, 159-177.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five models for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9*, 351-361.

Ghosh, B. K. (1979). A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association, 74*, 894-900.

Harvill, L. M. (1991). Standard errors of measurement. *Educational Measurement: Issues and Practice, 10,* 33-41.

Hogg, R. V., & Craig, A. T. (1995). *Introduction to mathematical statistics* (5th ed.). New Jersey: Prentice Hall.

Hogg, R. V., & Tanis, E. A. (1993). *Probability and statistical inference* (4th ed.). New York: Macmillan.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993a). *Iowa tests of basic skills* (Levels 9-14). Chicago: Riverside.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1993b). *Iowa tests of basic skills norms and score conversions*. Chicago: Riverside.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Jarjoura, D. (1985). Tolerance intervals for true scores. *Journal of Educational Statistics, 10*, 1-17.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 59-72.

Kolen, M. J., & Wang, T. (1998, April). *Conditional standard errors of measurement for composite scores using IRT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 285-307.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33,* 129-140.

Lee, W. (1998). *An investigation of procedures for estimating conditional scale-score standard errors of measurement and constructing confidence intervals for scale scores.* Unpublished doctoral dissertation, University of Iowa.

Lee, W., Brennan, R. L., & Kolen, M. J. (1998). *A comparison of some procedures for estimating conditional scale-score standard errors of measurement* (Iowa Testing Programs Occasional Paper No. 43). Iowa City, IA: University of Iowa.

Lee, W., & Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*, 1-20.

Leemis, L. M., & Trivedi, K. S. (1996). A comparison of approximate interval estimators for the Bernoulli parameter. *The American Statistician, 50*, 63-68.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325-336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement, 17*, 510-521.

Lord, F. M. (1984). Standard errors of measurement at different score levels. *Journal of Educational Measurement, 21*, 239-243.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453-461.

Mollenkopf, W. G. (1949). Variations of the standard error of measurement. *Psychometrika, 14*, 189-229.

Mooney, C. Z. (1997). *Monte Carlo simulation* (Sage University Paper series on Quantitative Applications in the Social Sciences, Series No. 07-116). Thousand Oaks, CA: Sage.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* McGraw-Hill.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-261). New York: American Council on Education and Macmillan.

Qualls-Payne, A. L. (1992). A comparison of score level estimates of the standard error of measurement. *Journal of Educational Measurement, 29*, 213-225.

Santner, T. J. (1998). A note on teaching large-sample binomial confidence intervals. *Teaching Statistics, 20*, 20-23.

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika, 41*, 275-278.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational measurement, 37*, 141-162.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22,* 209-212.

## TABLE 1

### Endpoints of Nominal 68% Intervals at Some Selected Observed Raw and DSS Scores with $k = 34$

| | $I_c(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ |
|---|---|---|---|---|---|---|
| $x = 5$, $DSS = 141$ | | | | | | |
| Raw | 2.9, 7.1 | 3.3, 7.4 | 3.2, 7.5 | 2.9, 8.0 | 3.6, 7.7 | |
| DSS | 133.4, 150.9 | 134.8, 152.3 | 134.6, 152.6 | 133.4, 154.8 | 136.0, 153.8 | 132.1, 149.9 |
| $x = 10$, $DSS = 165$ | | | | | | |
| Raw | 7.3, 12.7 | 7.6, 12.8 | 7.5, 12.9 | 7.1, 13.3 | 7.8, 13.0 | |
| DSS | 152.0, 173.8 | 153.2, 174.2 | 153.0, 174.4 | 151.2, 175.9 | 154.1, 174.8 | 153.8, 176.2 |
| $x = 17$, $DSS = 187$ | | | | | | |
| Raw | 14.1, 19.9 | 14.1, 19.9 | 14.1, 19.9 | 13.7, 20.3 | 14.2, 19.8 | |
| DSS | 178.2, 193.8 | 178.5, 193.6 | 178.3, 193.7 | 177.0, 194.8 | 178.6, 193.5 | 179.4, 194.6 |
| $x = 25$, $DSS = 207$ | | | | | | |
| Raw | 22.4, 27.6 | 22.2, 27.3 | 22.2, 27.4 | 21.7, 27.7 | 22.0, 27.1 | |
| DSS | 200.2, 218.1 | 199.8, 216.8 | 199.7, 217.0 | 198.4, 218.7 | 199.3, 215.8 | 198.1, 215.9 |
| $x = 30$, $DSS = 229$ | | | | | | |
| Raw | 28.1, 31.9 | 27.8, 31.5 | 27.7, 31.6 | 27.2, 31.9 | 27.4, 31.2 | |
| DSS | 220.4, 241.9 | 218.8, 239.2 | 218.5, 239.6 | 216.3, 241.8 | 217.1, 237.0 | 218.3, 239.7 |

**Raw-to-DSS Conversion**

| Raw | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DSS | 124 | 127 | 130 | 133 | 137 | 141 | 146 | 150 | 155 | 160 | 165 | 169 |
| Raw | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| DSS | 173 | 176 | 179 | 182 | 184 | 187 | 189 | 191 | 194 | 196 | 199 | 201 |
| Raw | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | |
| DSS | 204 | 207 | 210 | 214 | 219 | 224 | 229 | 234 | 241 | 250 | 261 | |

## TABLE 2

### Average Actual Coverage Probabilities of Nominal 95% Intervals with $k = 34$

| | | $I_c(\tau)$ | $I_o(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ | $I_o(\xi)$ |
|---|---|---|---|---|---|---|---|---|---|
| Raw | Mean | .924 | .947 | .952 | .967 | .970 | .953 | | |
| | SD | .062 | .028 | .012 | .010 | .009 | .014 | | |
| DSS | Mean | .924 | .947 | .952 | .967 | .970 | .953 | .944 | .949 |
| | SD | .062 | .028 | .012 | .010 | .009 | .014 | .068 | .032 |
| GE | Mean | .924 | .947 | .952 | .967 | .970 | .953 | .944 | .949 |
| | SD | .062 | .028 | .012 | .010 | .009 | .014 | .070 | .053 |
| PR | Mean | .924 | .947 | .952 | .967 | .970 | .953 | .889 | .941 |
| | SD | .062 | .028 | .012 | .010 | .009 | .014 | .074 | .037 |
| ST | Mean | .924 | .947 | .952 | .967 | .970 | .953 | .911 | .943 |
| | SD | .062 | .028 | .012 | .010 | .009 | .014 | .087 | .039 |

Note: Shaded area is for $I_v(\xi)$ = scale-score counterparts using true-score conversions;

$I_c(\tau)$ = conditional confidence intervals using conditional SEMs;

$I_o(\tau)$ = overall confidence intervals using overall SEMs;

$I_s(\tau)$ = score confidence intervals;

$I_b(\tau)$ = Bayes confidence intervals;

$I_e(\tau)$ = Clopper-Pearson exact confidence intervals;

$I_p(\tau)$ = credibility intervals with non-informative prior;

$I_c(\xi)$ = conditional scale-score confidence intervals using CSSEMs; and

$I_o(\xi)$ = overall scale-score confidence intervals using overall scale-score SEMs.

## TABLE 3

**Average Actual Coverage Probabilities of Nominal 68% Intervals with $k = 34$**

| | | $I_c(\tau)$ | $I_o(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ | $I_o(\xi)$ |
|---|---|---|---|---|---|---|---|---|---|
| Raw | Mean | .676 | .693 | .681 | .691 | .766 | .686 | | |
| | SD | .058 | .093 | .047 | .047 | .046 | .052 | | |
| DSS | Mean | .676 | .693 | .681 | .691 | .766 | .686 | .697 | .709 |
| | SD | .058 | .093 | .047 | .047 | .046 | .052 | .071 | .096 |
| GE | Mean | .676 | .693 | .681 | .691 | .766 | .686 | .700 | .728 |
| | SD | .058 | .093 | .047 | .047 | .046 | .052 | .083 | .126 |
| PR | Mean | .676 | .693 | .681 | .691 | .766 | .686 | .652 | .695 |
| | SD | .058 | .093 | .047 | .047 | .046 | .052 | .066 | .144 |
| ST | Mean | .676 | .693 | .681 | .691 | .766 | .686 | .670 | .701 |
| | SD | .058 | .093 | .047 | .047 | .046 | .052 | .153 | .194 |

Note: Shaded area is for $I_v(\xi)$ = scale-score counterparts using true-score conversions;
$I_c(\tau)$ = conditional confidence intervals using conditional SEMs;
$I_o(\tau)$ = overall confidence intervals using overall SEMs;
$I_s(\tau)$ = score confidence intervals;
$I_b(\tau)$ = Bayes confidence intervals;
$I_e(\tau)$ = Clopper-Pearson exact confidence intervals;
$I_p(\tau)$ = credibility intervals with non-informative prior;
$I_c(\xi)$ = conditional scale-score confidence intervals using CSSEMs; and
$I_o(\xi)$ = overall scale-score confidence intervals using overall scale-score SEMs.

## TABLE 4

### Average Actual Coverage Probabilities of Nominal 50% Intervals with $k = 34$

| | | $I_c(\tau)$ | $I_o(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ | $I_o(\xi)$ |
|---|---|---|---|---|---|---|---|---|---|
| Raw | Mean | .497 | .519 | .499 | .503 | .617 | .501 | | |
| | SD | .064 | .114 | .060 | .058 | .057 | .063 | | |
| DSS | Mean | .497 | .519 | .499 | .503 | .617 | .501 | 516 | .526 |
| | SD | .064 | .114 | .060 | .058 | .057 | .063 | 074 | .102 |
| GE | Mean | .497 | .519 | .499 | .503 | .617 | .501 | 525 | .545 |
| | SD | .064 | .114 | .060 | .058 | .057 | .063 | 091 | .130 |
| PR | Mean | .497 | .519 | .499 | .503 | .617 | .501 | 487 | .537 |
| | SD | .064 | .114 | .060 | .058 | .057 | .063 | 074 | .188 |
| ST | Mean | .497 | .519 | .499 | .503 | .617 | .501 | 490 | .521 |
| | SD | .064 | .114 | .060 | .058 | .057 | .063 | 155 | .164 |

Note: Shaded area is for $I_v(\xi)$ = scale-score counterparts using true-score conversions;

$I_c(\tau)$ = conditional confidence intervals using conditional SEMs;

$I_o(\tau)$ = overall confidence intervals using overall SEMs;

$I_s(\tau)$ = score confidence intervals;

$I_b(\tau)$ = Bayes confidence intervals;

$I_e(\tau)$ = Clopper-Pearson exact confidence intervals;

$I_p(\tau)$ = credibility intervals with non-informative prior;

$I_c(\xi)$ = conditional scale-score confidence intervals using CSSEMs; and

$I_o(\xi)$ = overall scale-score confidence intervals using overall scale-score SEMs.

## TABLE 5

### Average Actual Coverage Probabilities of Nominal 95% Intervals with $k = 17$

| | | $I_c(\tau)$ | $I_o(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ | $I_o(\xi)$ |
|---|---|---|---|---|---|---|---|---|---|
| Raw | Mean | .898 | .947 | .954 | .977 | .976 | .954 | | |
| | SD | .089 | .030 | .015 | .010 | .010 | .034 | | |
| DSS | Mean | .898 | .947 | .954 | .977 | .976 | .954 | .928 | .951 |
| | SD | .089 | .030 | .015 | .010 | .010 | .034 | .103 | .029 |
| GE | Mean | .898 | .947 | .954 | .977 | .976 | .954 | .926 | .947 |
| | SD | .089 | .030 | .015 | .010 | .010 | .034 | .105 | .060 |
| PR | Mean | .898 | .947 | .954 | .977 | .976 | .954 | .856 | .940 |
| | SD | .089 | .030 | .015 | .010 | .010 | .034 | .090 | .043 |
| ST | Mean | .898 | .947 | .954 | .977 | .976 | .954 | .887 | .947 |
| | SD | .089 | .030 | .015 | .010 | .010 | .034 | .104 | .031 |

Note: Shaded area is for $I_v(\xi)$ = scale-score counterparts using true-score conversions;
$I_c(\tau)$ = conditional confidence intervals using conditional SEMs;
$I_o(\tau)$ = overall confidence intervals using overall SEMs;
$I_s(\tau)$ = score confidence intervals;
$I_b(\tau)$ = Bayes confidence intervals;
$I_e(\tau)$ = Clopper-Pearson exact confidence intervals;
$I_p(\tau)$ = credibility intervals with non-informative prior;
$I_c(\xi)$ = conditional scale-score confidence intervals using CSSEMs; and
$I_o(\xi)$ = overall scale-score confidence intervals using overall scale-score SEMs.

## TABLE 6

### Average Actual Coverage Probabilities of Nominal 68% Intervals with $k = 17$

| | | $I_c(\tau)$ | $I_o(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ | $I_o(\xi)$ |
|---|---|---|---|---|---|---|---|---|---|
| Raw | Mean | .666 | .693 | .676 | .697 | .803 | .690 | | |
| | SD | .075 | .115 | .063 | .063 | .052 | .062 | | |
| DSS | Mean | .666 | .693 | .676 | .697 | .803 | .690 | .703 | .707 |
| | SD | .075 | .115 | .063 | .063 | .052 | .062 | .106 | .104 |
| GE | Mean | .666 | .693 | .676 | .697 | .803 | .690 | .709 | .731 |
| | SD | .075 | .115 | .063 | .063 | .052 | .062 | .108 | .129 |
| PR | Mean | .666 | .693 | .676 | .697 | .803 | .690 | .635 | .695 |
| | SD | .075 | .115 | .063 | .063 | .052 | .062 | .079 | .149 |
| ST | Mean | .666 | .693 | .676 | .697 | .803 | .690 | .665 | .702 |
| | SD | .075 | .115 | .063 | .063 | .052 | .062 | .136 | .138 |

Note: Shaded area is for $I_v(\xi)$ = scale-score counterparts using true-score conversions;
$I_c(\tau)$ = conditional confidence intervals using conditional SEMs;
$I_o(\tau)$ = overall confidence intervals using overall SEMs;
$I_s(\tau)$ = score confidence intervals;
$I_b(\tau)$ = Bayes confidence intervals;
$I_e(\tau)$ = Clopper-Pearson exact confidence intervals;
$I_p(\tau)$ = credibility intervals with non-informative prior;
$I_c(\xi)$ = conditional scale-score confidence intervals using CSSEMs; and
$I_o(\xi)$ = overall scale-score confidence intervals using overall scale-score SEMs.

## TABLE 7

### Average Actual Coverage Probabilities of Nominal 50% Intervals with $k = 17$

| | | $I_c(\tau)$ | $I_o(\tau)$ | $I_s(\tau)$ | $I_b(\tau)$ | $I_e(\tau)$ | $I_p(\tau)$ | $I_c(\xi)$ | $I_o(\xi)$ |
|---|---|---|---|---|---|---|---|---|---|
| Raw | Mean | .495 | .513 | .496 | .505 | .661 | .510 | | |
| | SD | .088 | .130 | .090 | .086 | .078 | .084 | | |
| DSS | Mean | .495 | .513 | .496 | .505 | .661 | .510 | .530 | .531 |
| | SD | .088 | .130 | .090 | .086 | .078 | .084 | .099 | .118 |
| GE | Mean | .495 | .513 | .496 | .505 | .661 | .510 | .535 | .552 |
| | SD | .088 | .130 | .090 | .086 | .078 | .084 | .108 | .149 |
| PR | Mean | .495 | .513 | .496 | .505 | .661 | .510 | .468 | .526 |
| | SD | .088 | .130 | .090 | .086 | .078 | .084 | .082 | .192 |
| ST | Mean | .495 | .513 | .496 | .505 | .661 | .510 | .484 | .526 |
| | SD | .088 | .130 | .090 | .086 | .078 | .084 | .178 | .209 |

Note: Shaded area is for $I_v(\xi)$ = scale-score counterparts using true-score conversions;
$I_c(\tau)$ = conditional confidence intervals using conditional SEMs;
$I_o(\tau)$ = overall confidence intervals using overall SEMs;
$I_s(\tau)$ = score confidence intervals;
$I_b(\tau)$ = Bayes confidence intervals;
$I_e(\tau)$ = Clopper-Pearson exact confidence intervals;
$I_p(\tau)$ = credibility intervals with non-informative prior;
$I_c(\xi)$ = conditional scale-score confidence intervals using CSSEMs; and
$I_o(\xi)$ = overall scale-score confidence intervals using overall scale-score SEMs.

***FIGURE 1*. True-Score and Observed-Score Conversions**

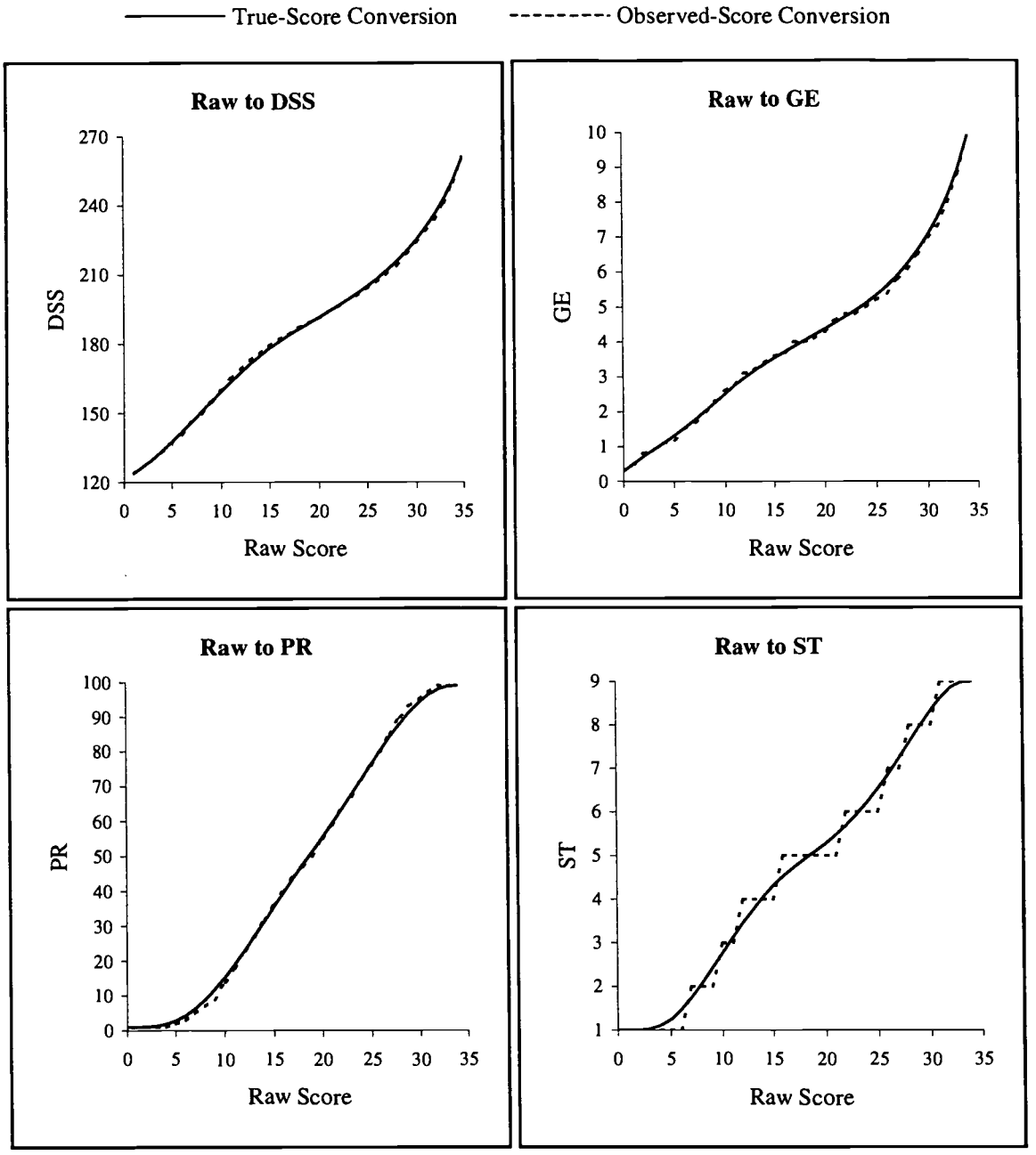


*FIGURE 2* . **Lengths of Nominal 68% Intervals for Raw and DSS Scores with $k = 34$**

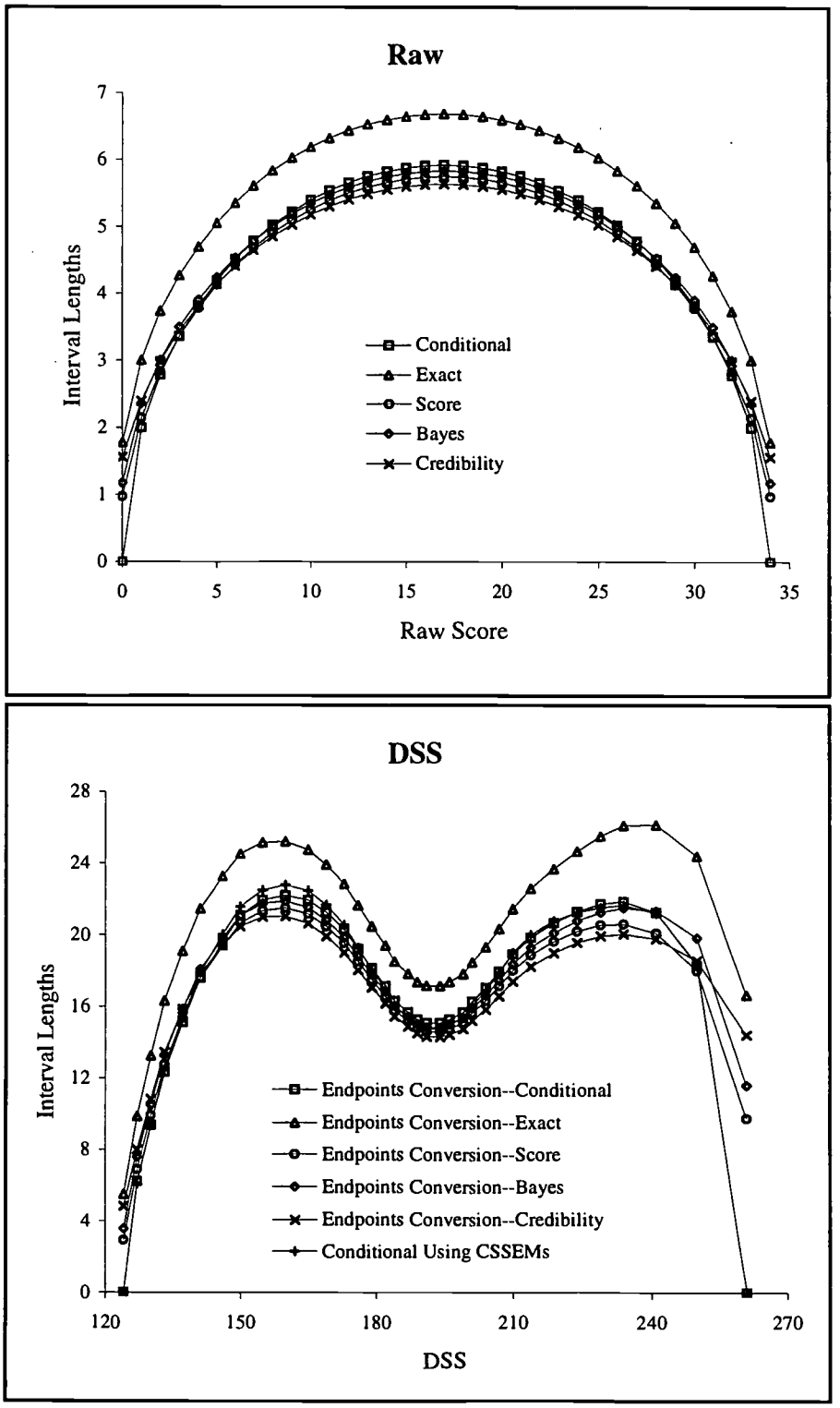


*FIGURE 3*. **Raw to Scale-Score Transformations with $k = 34$**

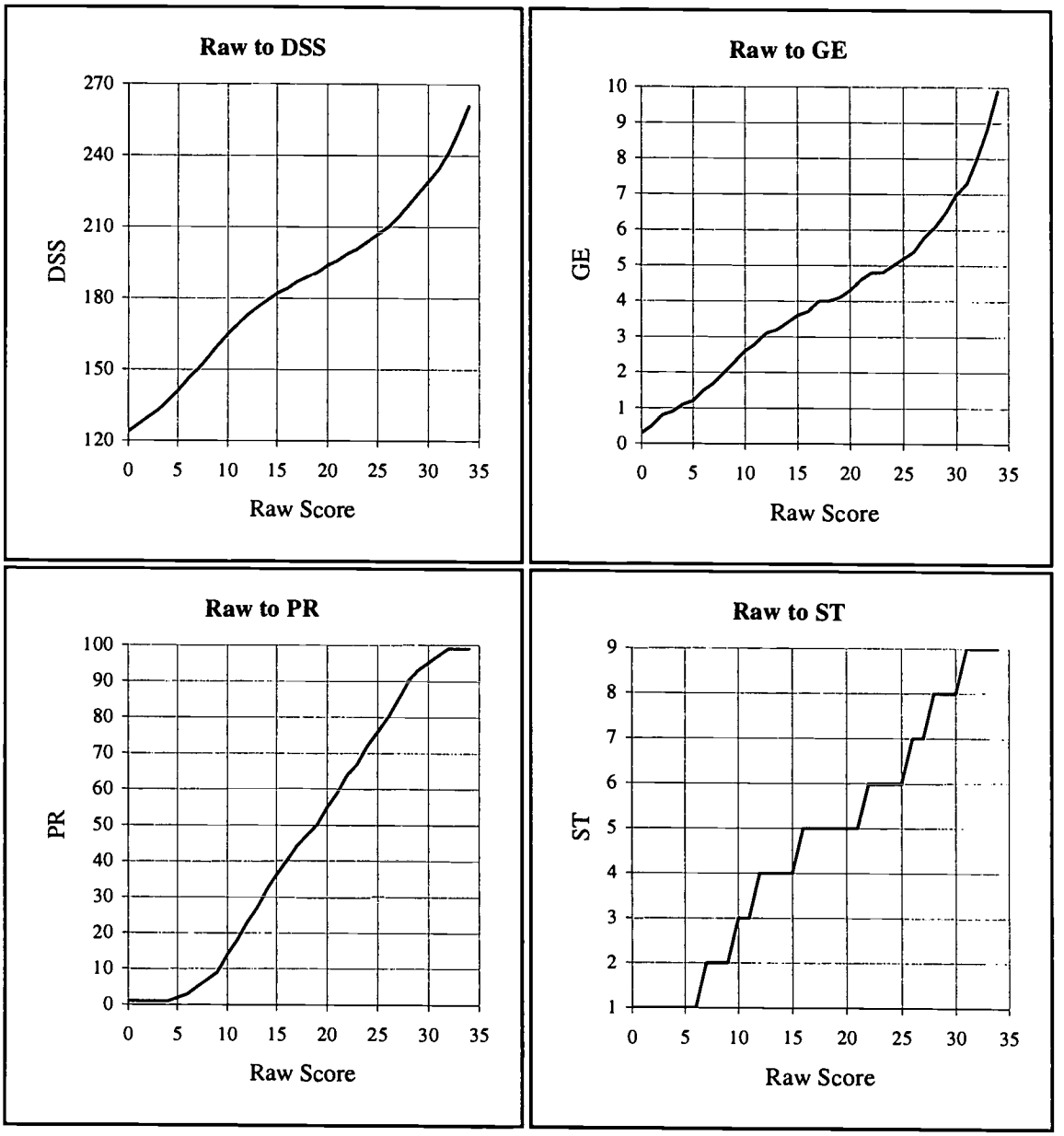


*FIGURE 4*. **Raw to Scale-Score Transformations with $k = 17$**

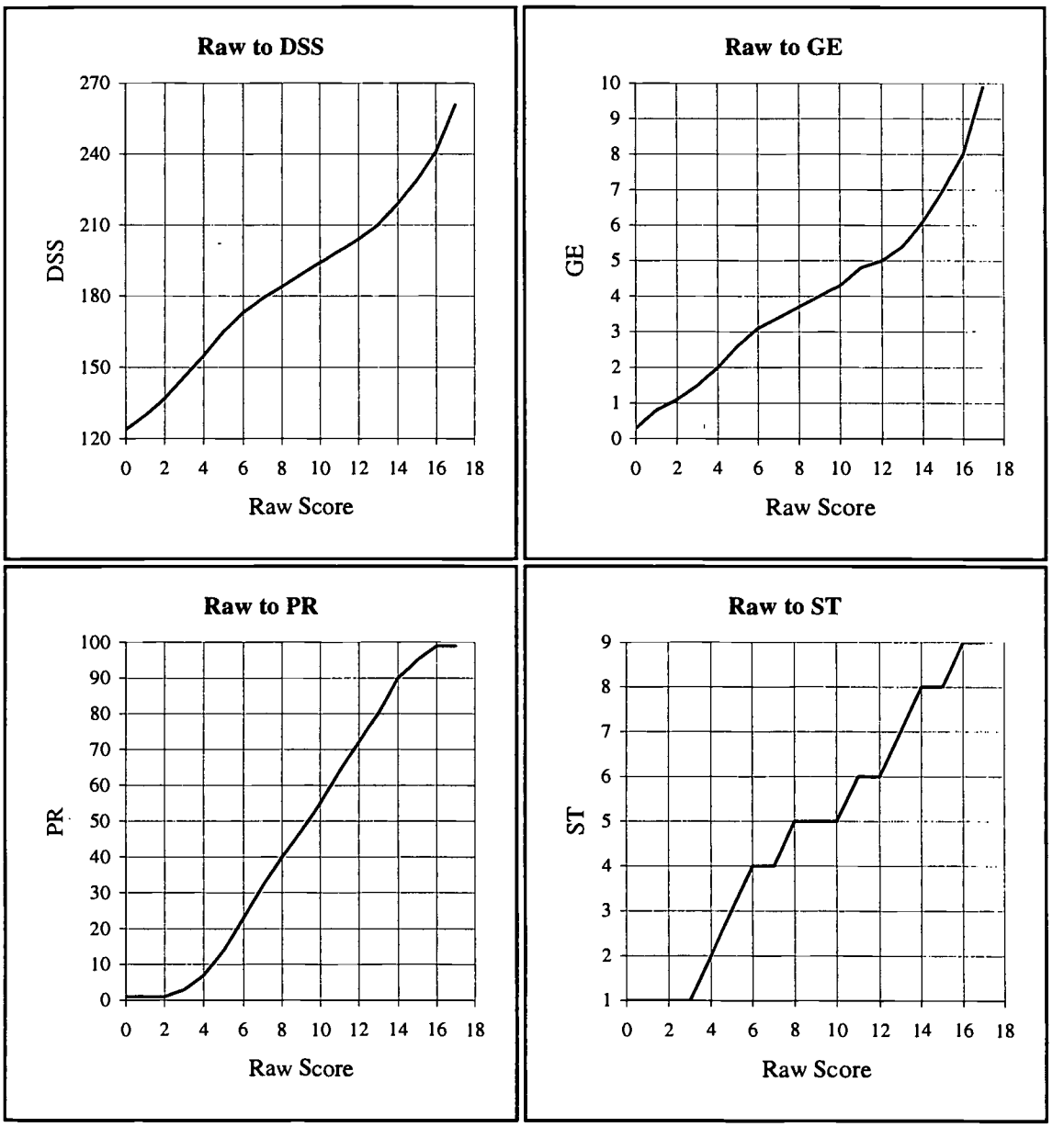


*FIGURE 5*. **Actual Coverage Probabilities of Nominal 95% Raw-Score Intervals with $k$ = 34**

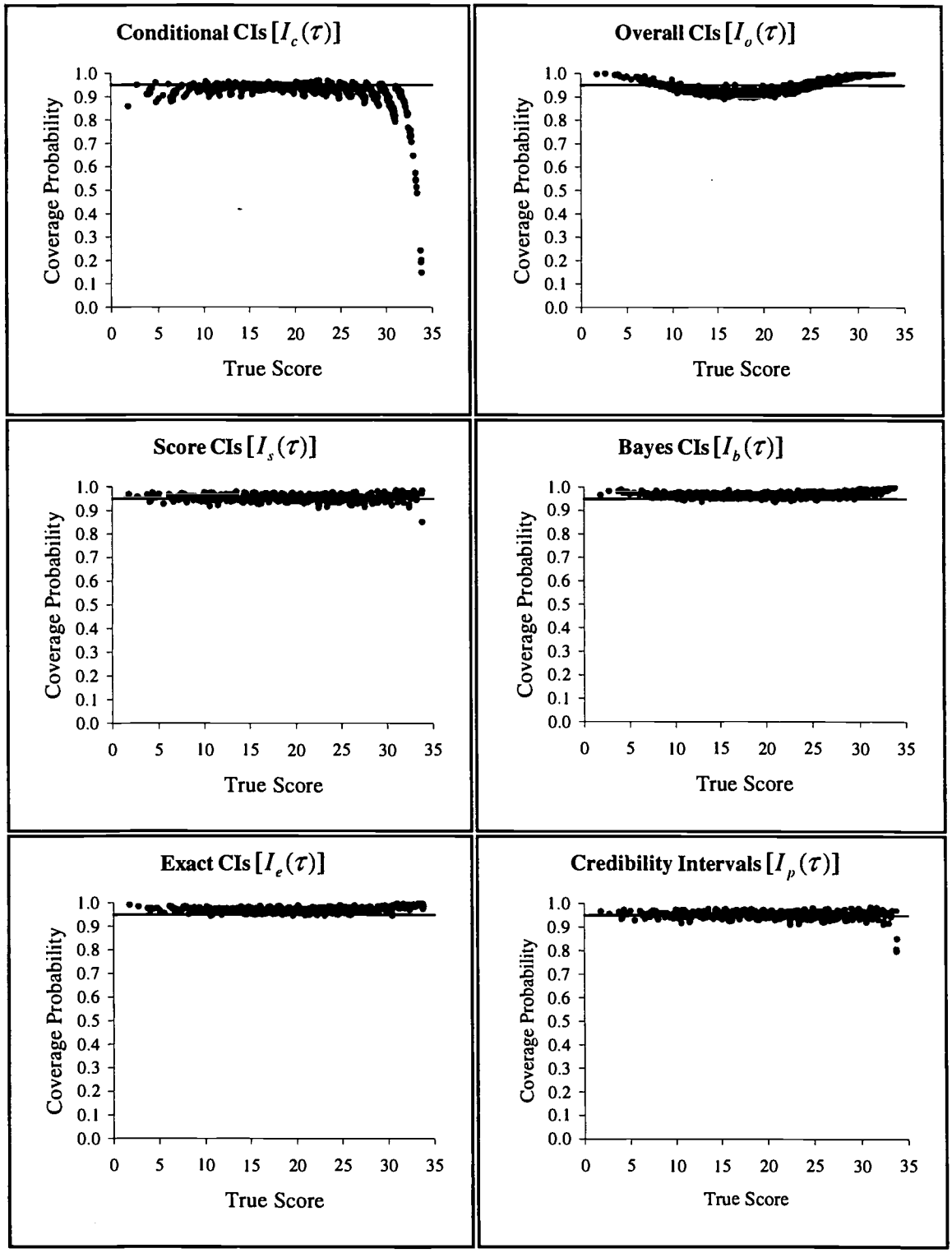


*FIGURE 6.* **Actual Coverage Probabilities of Nominal 95% Scale - Score Intervals Using Conditional Scale - Score SEMs $[I_c(\xi)]$ with $k = 34$**

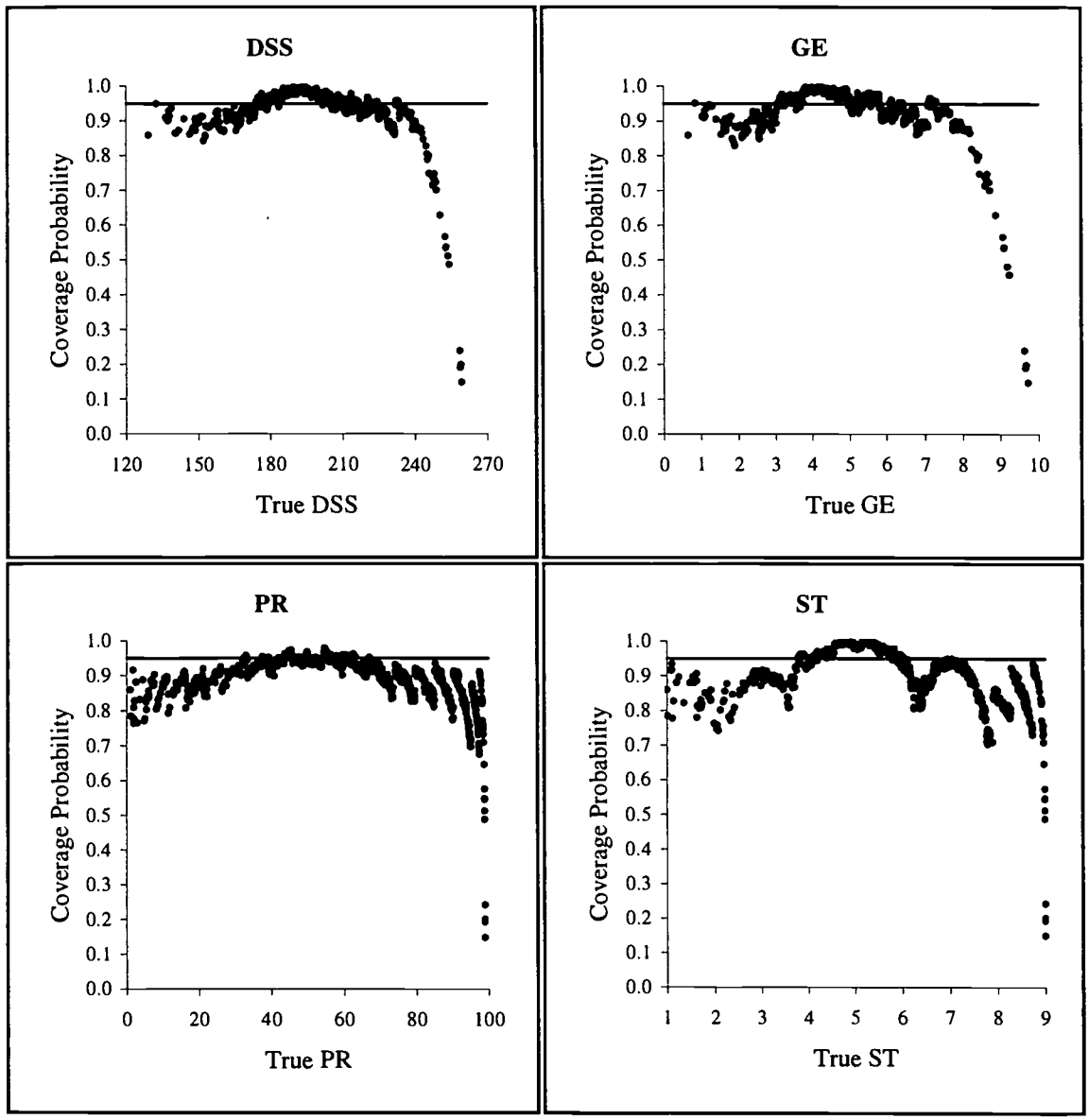


*FIGURE 7.* **Actual Coverage Probabilities of Nominal 95% Scale - Score Intervals Using Overall Scale - Score SEMs $[I_o(\xi)]$ with $k = 34$**

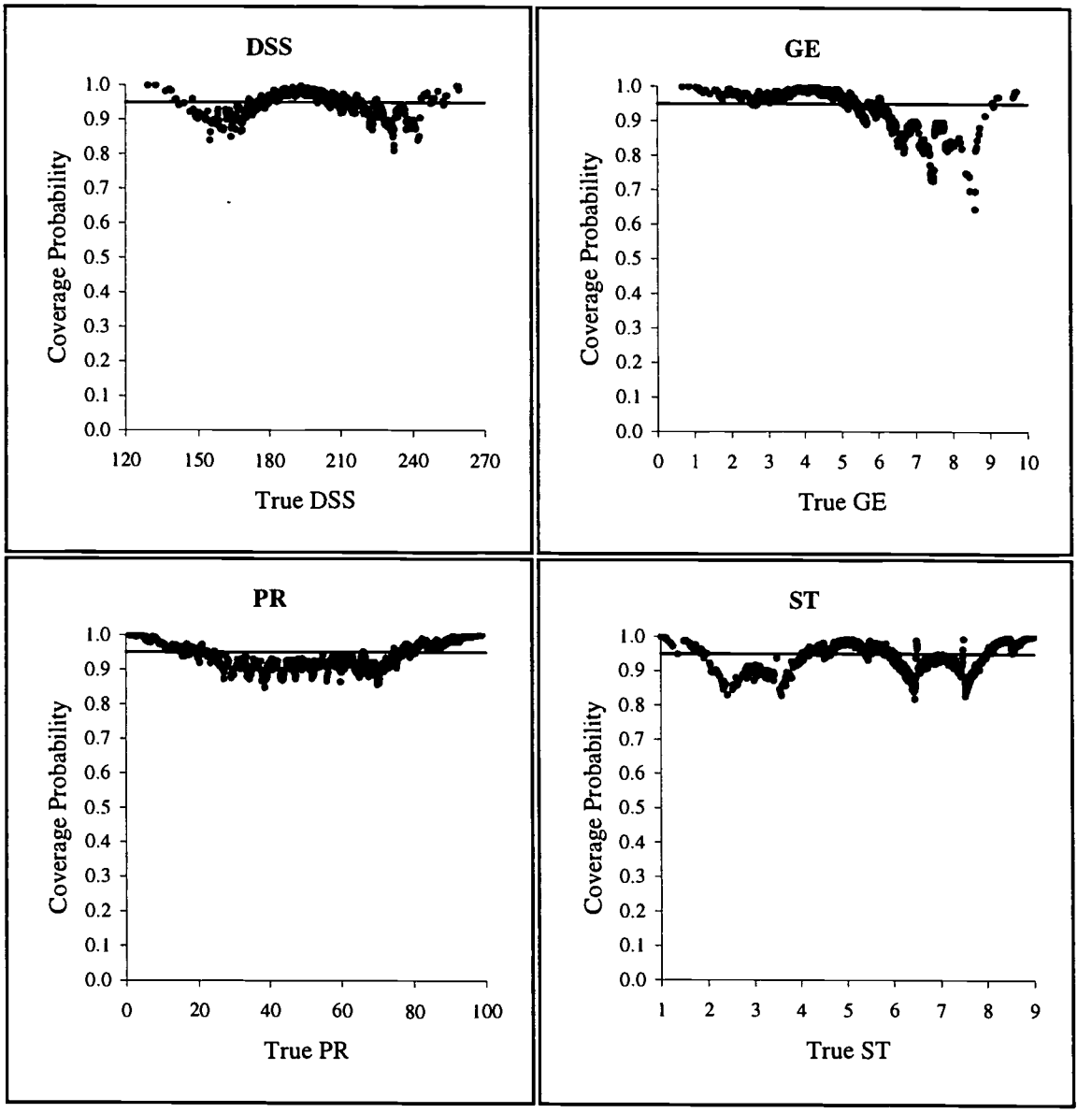


*FIGURE 8*. **True and Estimated SEMs with $k$ = 34**

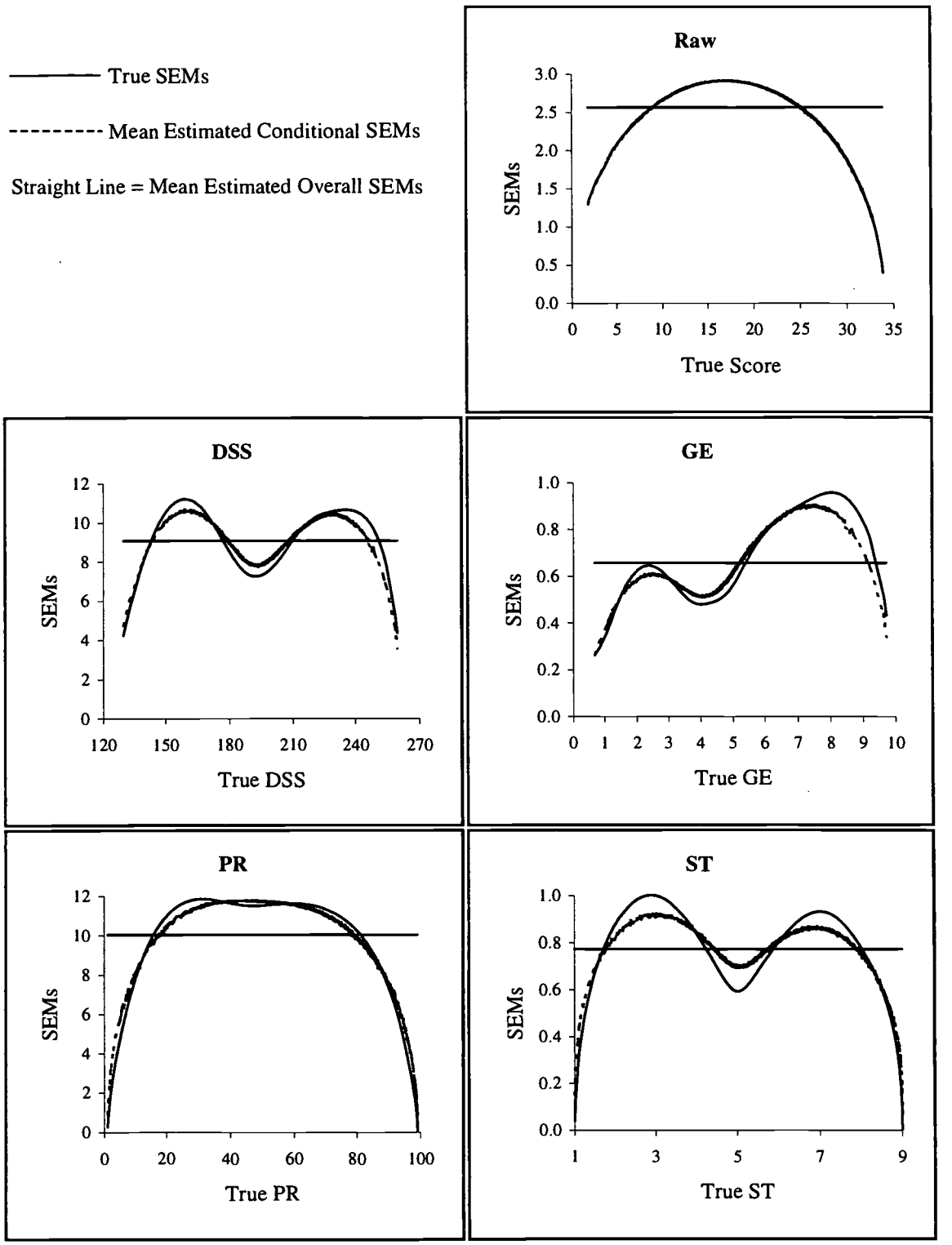


*FIGURE 9*. **Bias for Estimated SEMs with $k = 34$**

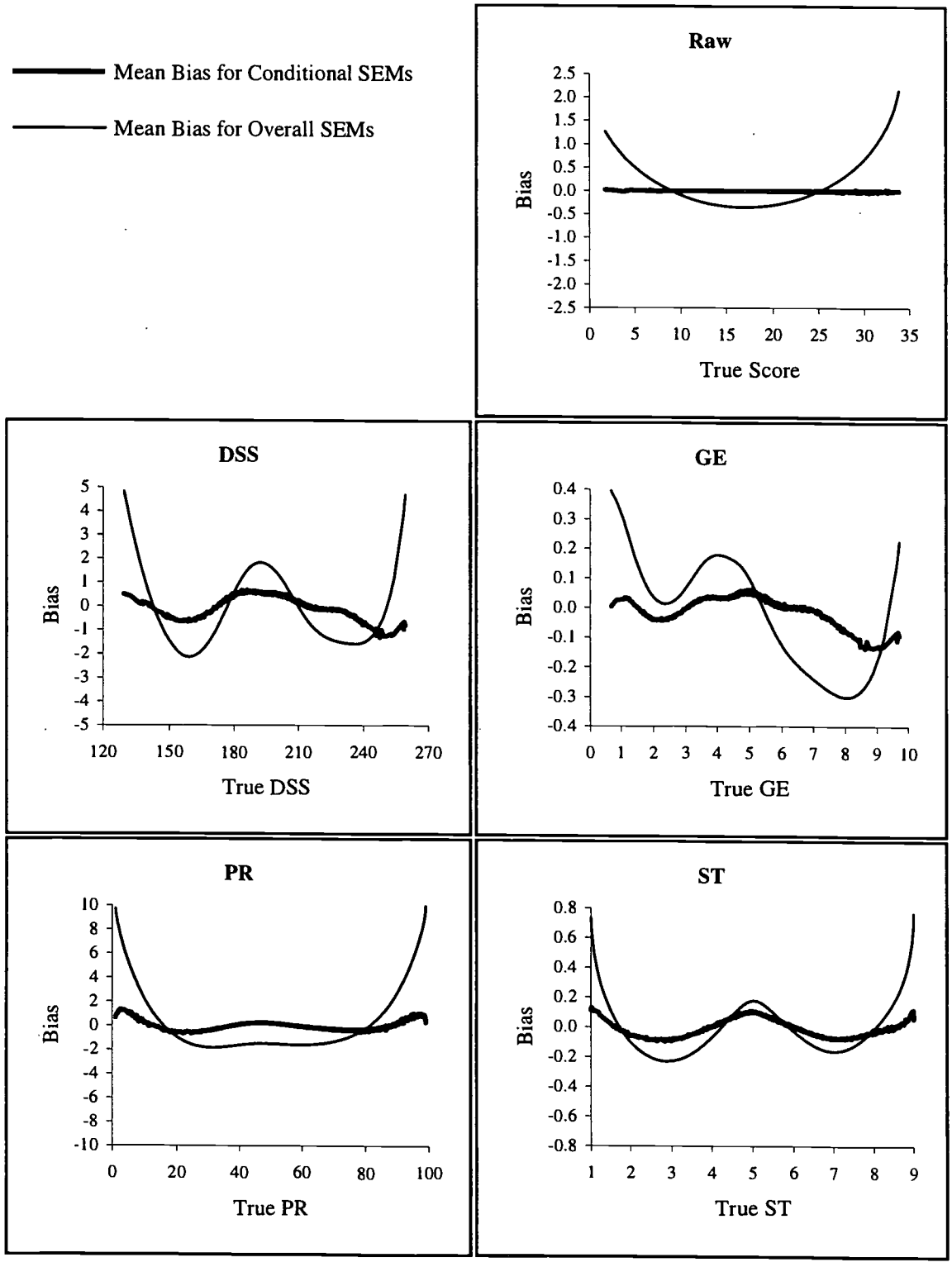


***FIGURE 10*. Actual Coverage Probabilities of Nominal 95% Conditional Confidence Intervals Using True Conditional SEMs with $k$ = 34**

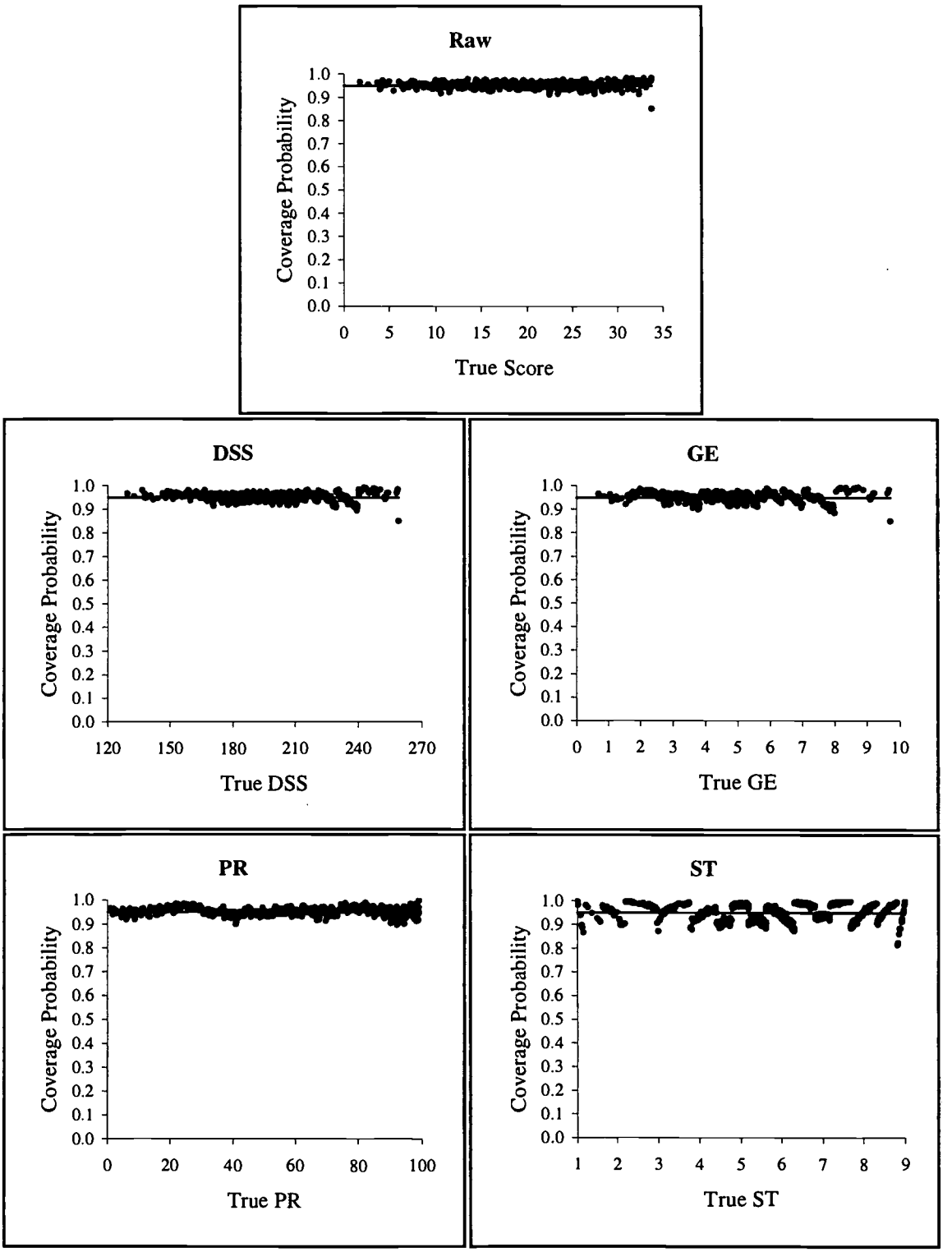


***FIGURE 11*. Actual Coverage Probabilities of Nominal 68% Raw-Score Intervals with $k = 34$**

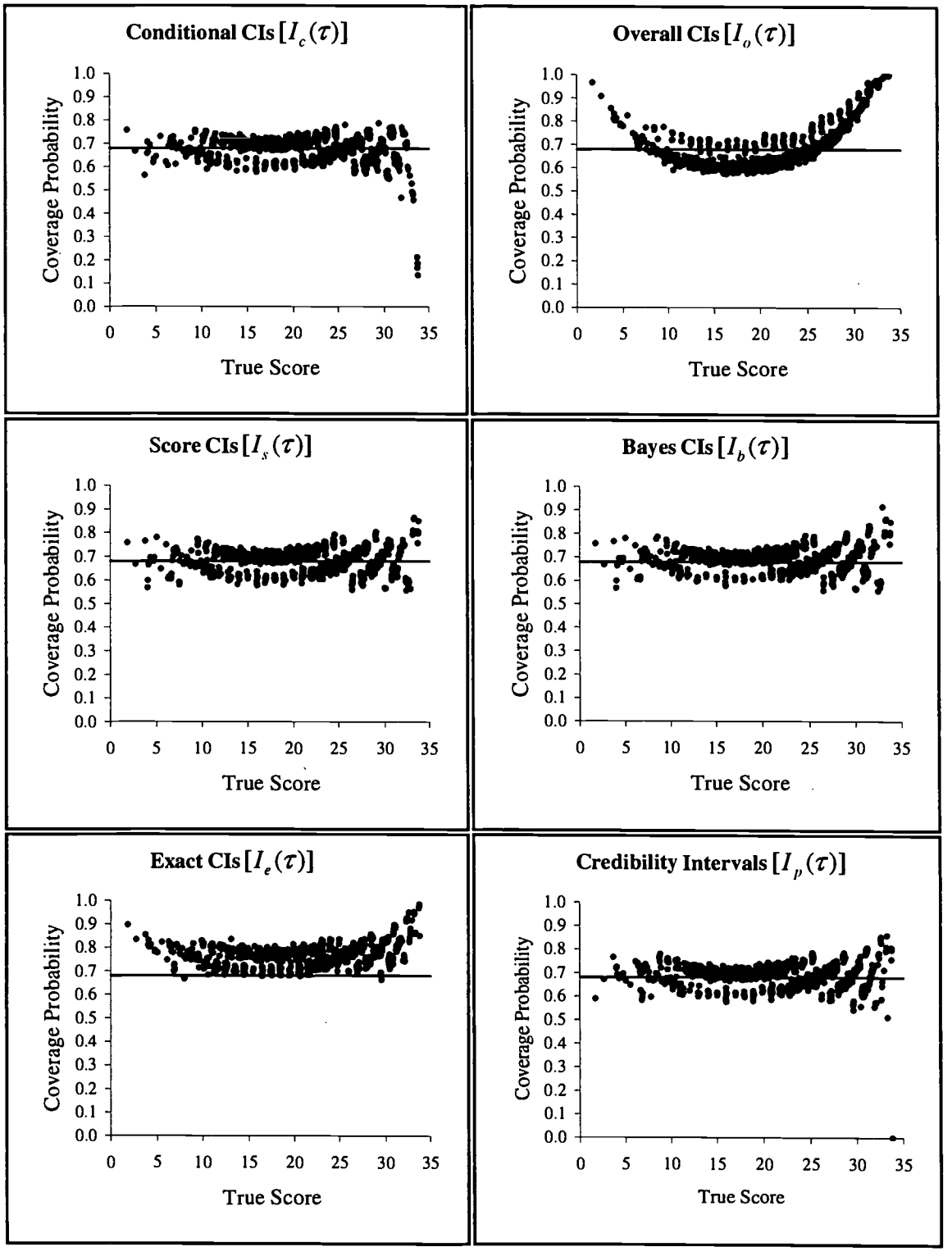


***FIGURE 12.*** **Actual Coverage Probabilities of Nominal 68% Scale - Score Intervals Using Conditional Scale - Score SEMs $[I_c(\xi)]$ with $k = 34$**

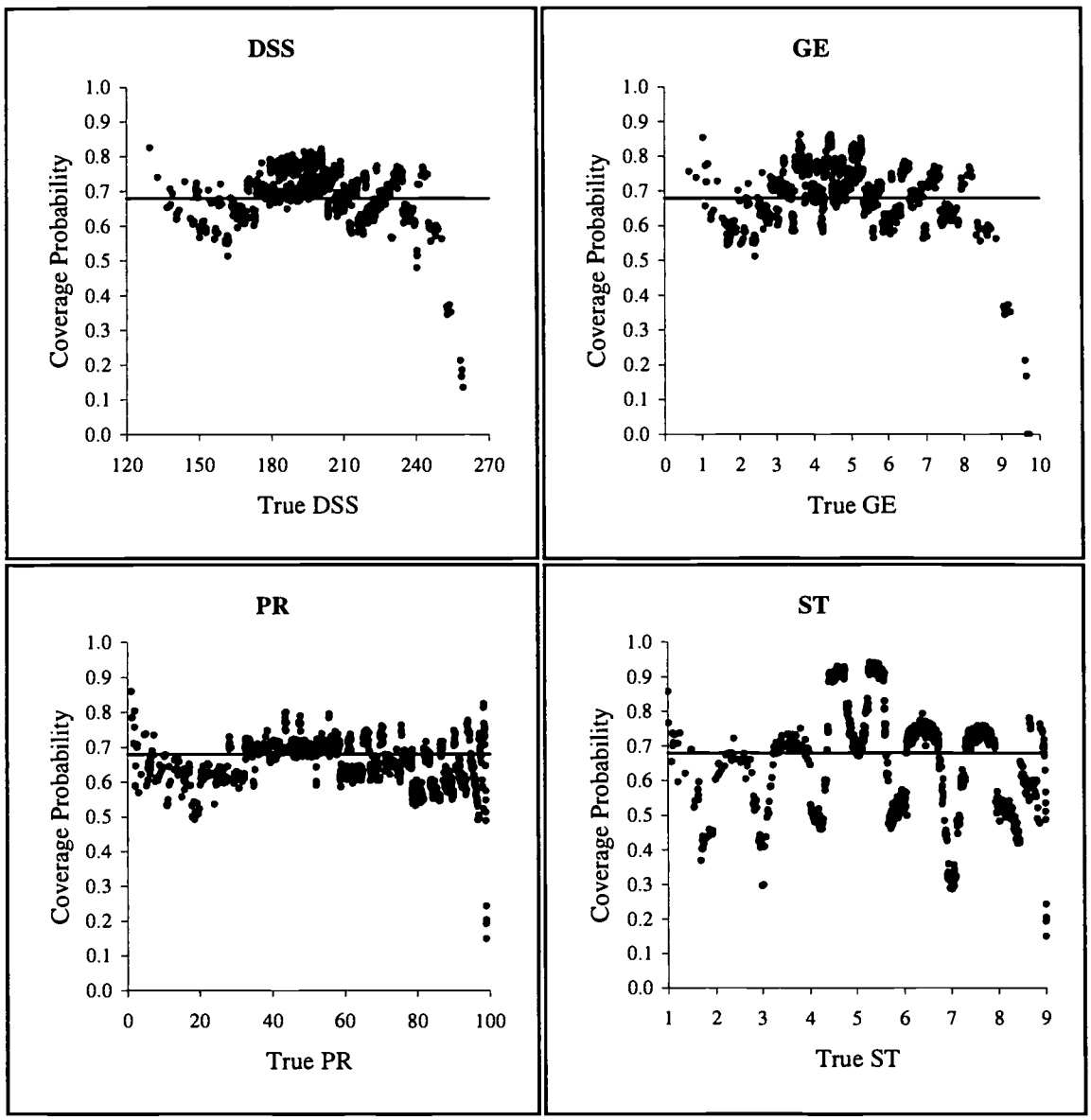


*FIGURE 13.* **Actual Coverage Probabilities of Nominal 68% Scale - Score Intervals Using Overall Scale - Score SEMs $[I_o(\xi)]$ with $k = 34$**

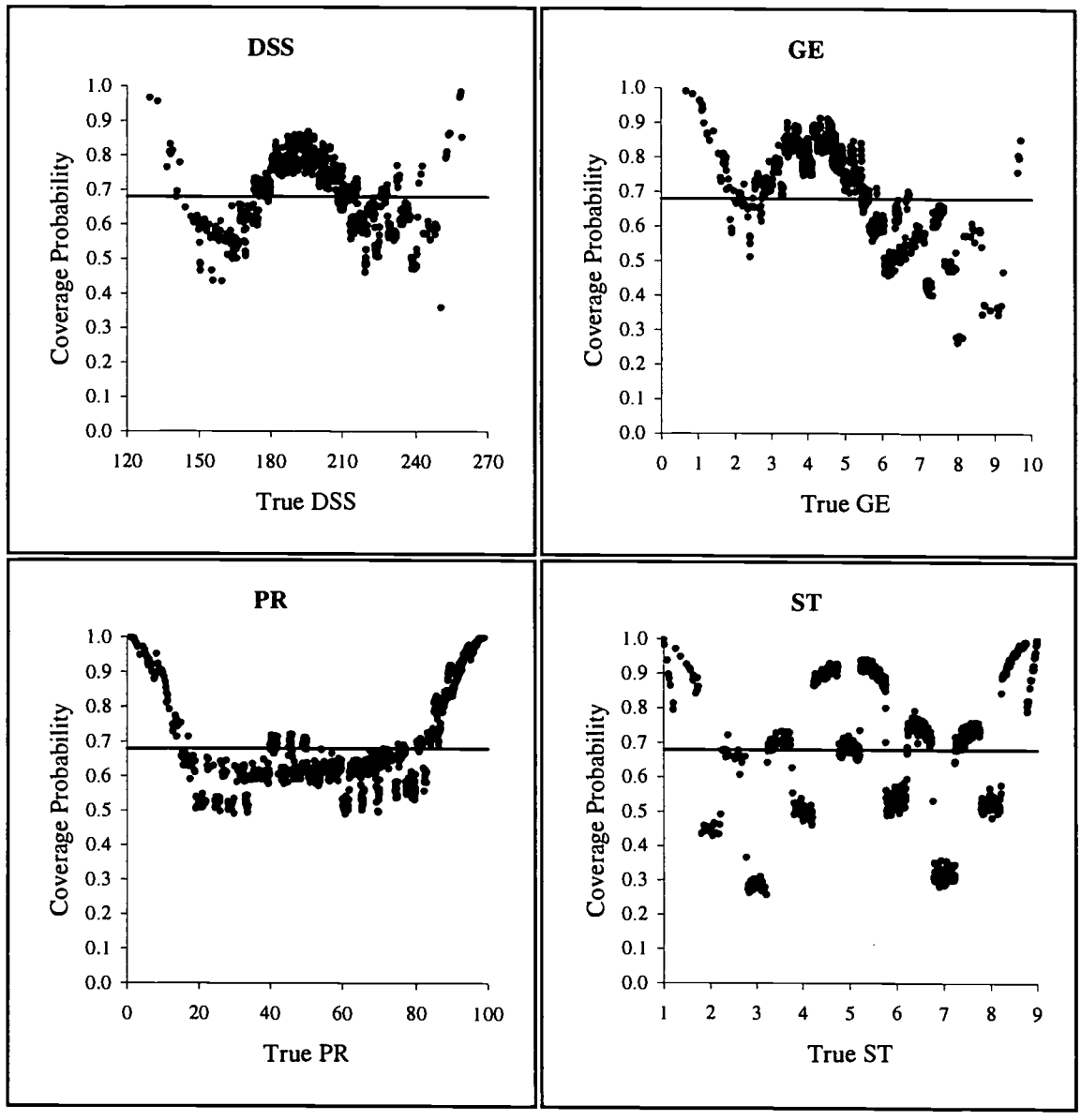


*FIGURE 14*. **Actual Coverage Probabilities of Nominal 68% Conditional Confidence Intervals Using True Conditional SEMs with $k$ = 34**

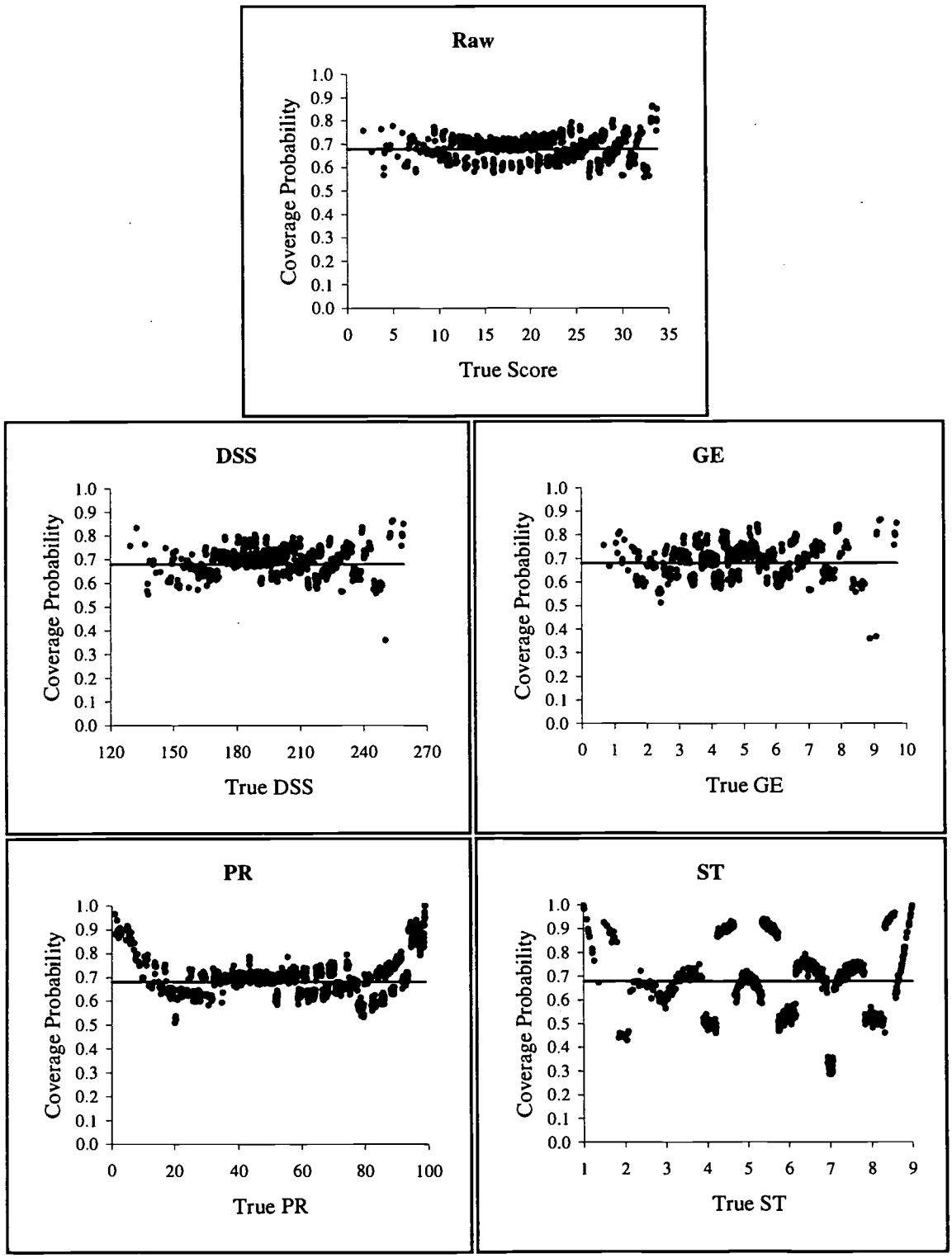


*FIGURE 15* . **Actual Coverage Probabilities of Nominal 50% Raw-Score Intervals with $k$ = 34**

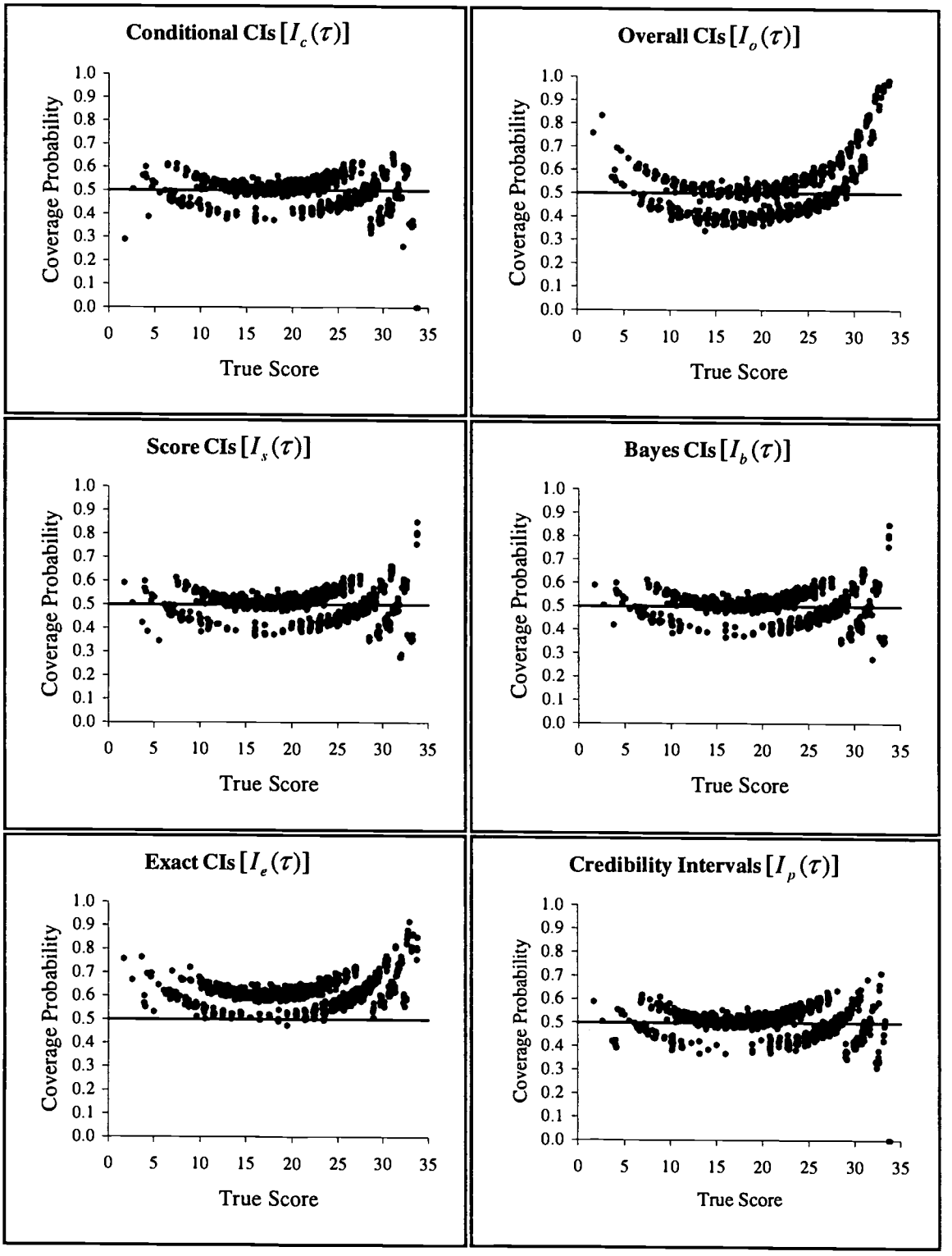


*FIGURE 16.* **Actual Coverage Probabilities of Nominal 50% Scale - Score Intervals Using Conditional Scale - Score SEMs $[I_c(\xi)]$ with $k = 34$**

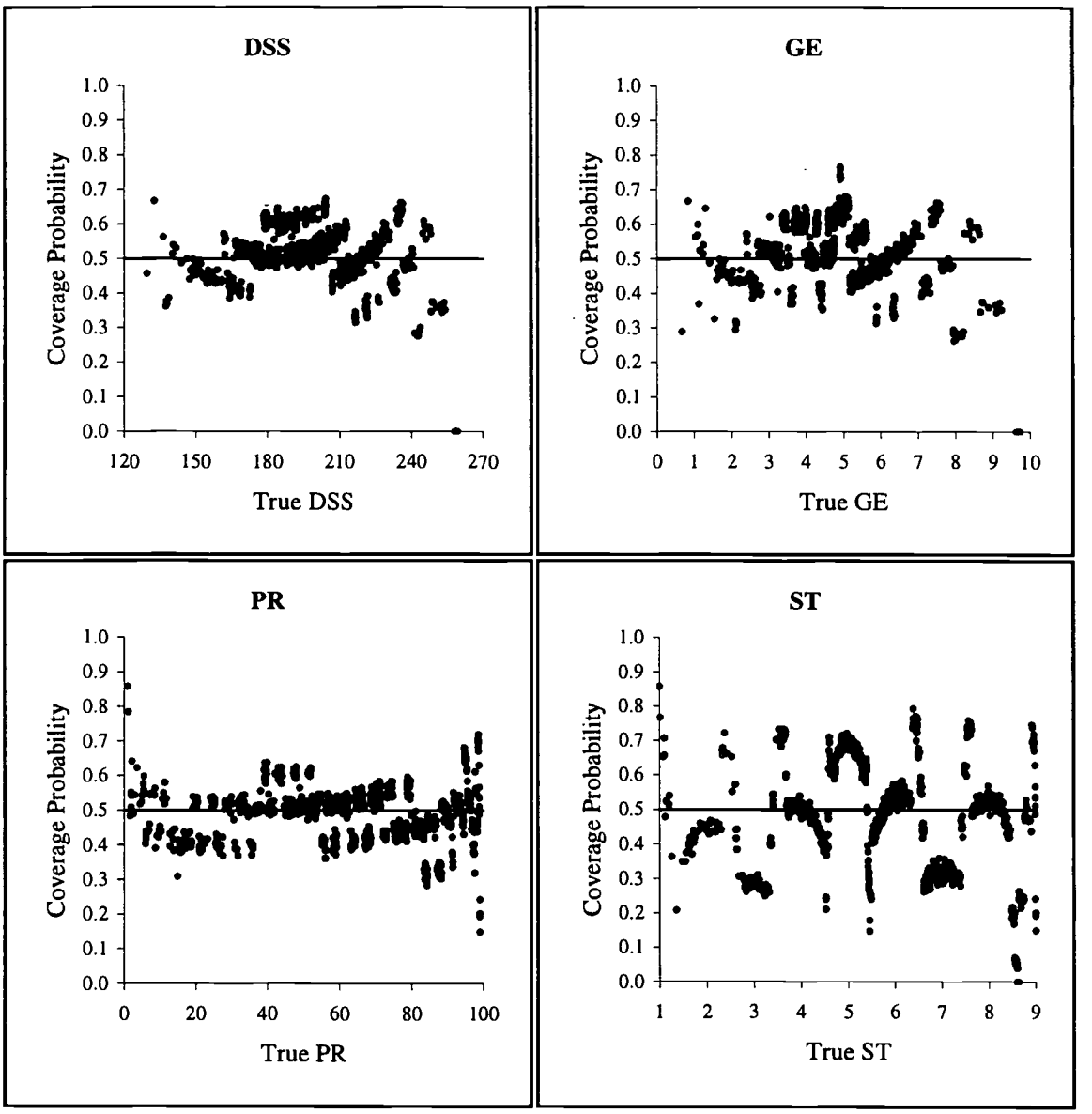


*FIGURE 17.* **Actual Coverage Probabilities of Nominal 50% Scale - Score Intervals Using Overall Scale - Score SEMs $[I_o(\xi)]$ with $k = 34$**

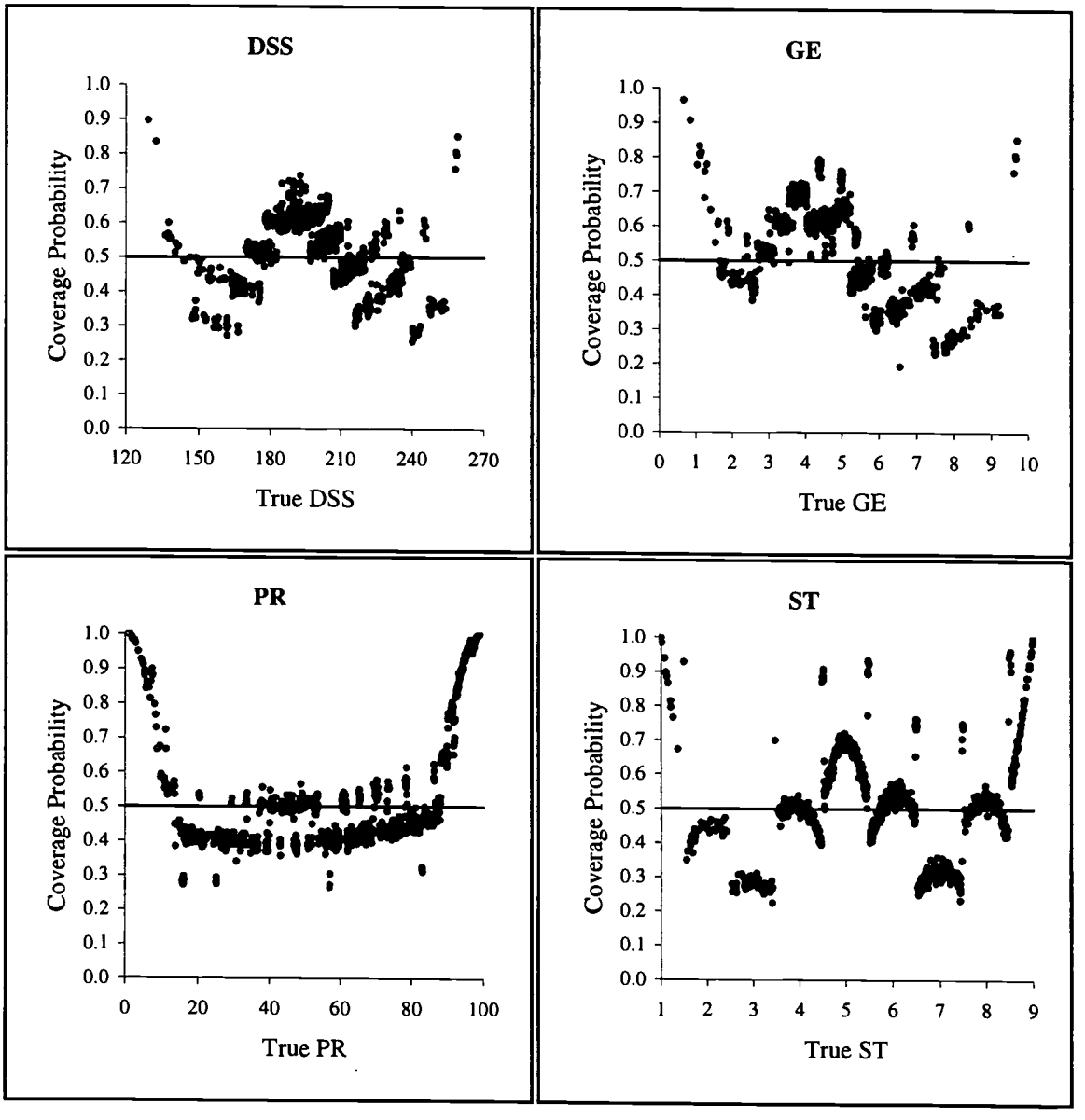


*FIGURE 18*. **Actual Coverage Probabilities of Nominal 50% Conditional Confidence Intervals Using True Conditional SEMs with $k$ = 34**

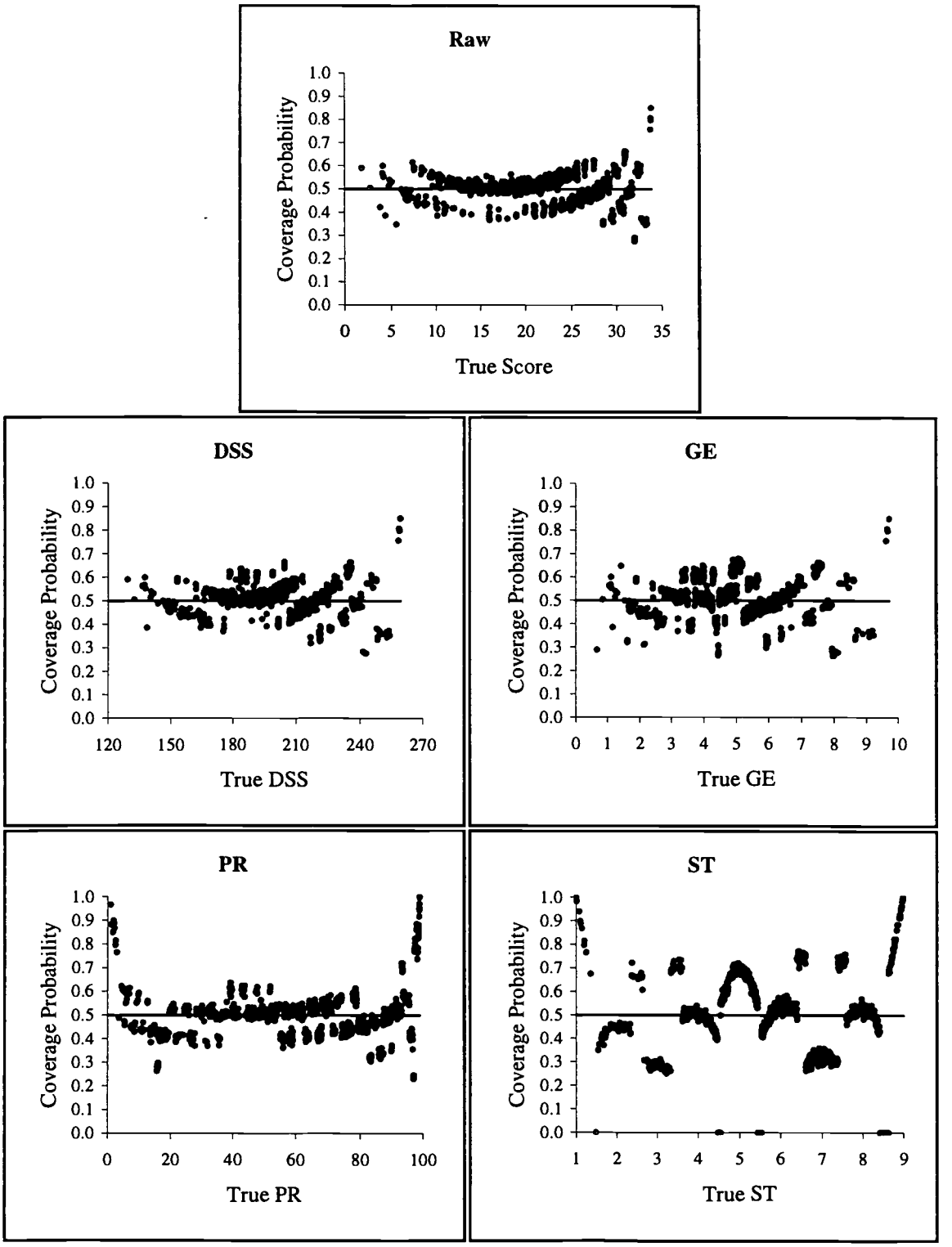


*FIGURE 19*. **Actual Coverage Probabilities of Nominal 95% Raw-Score Intervals with $k$ = 17**

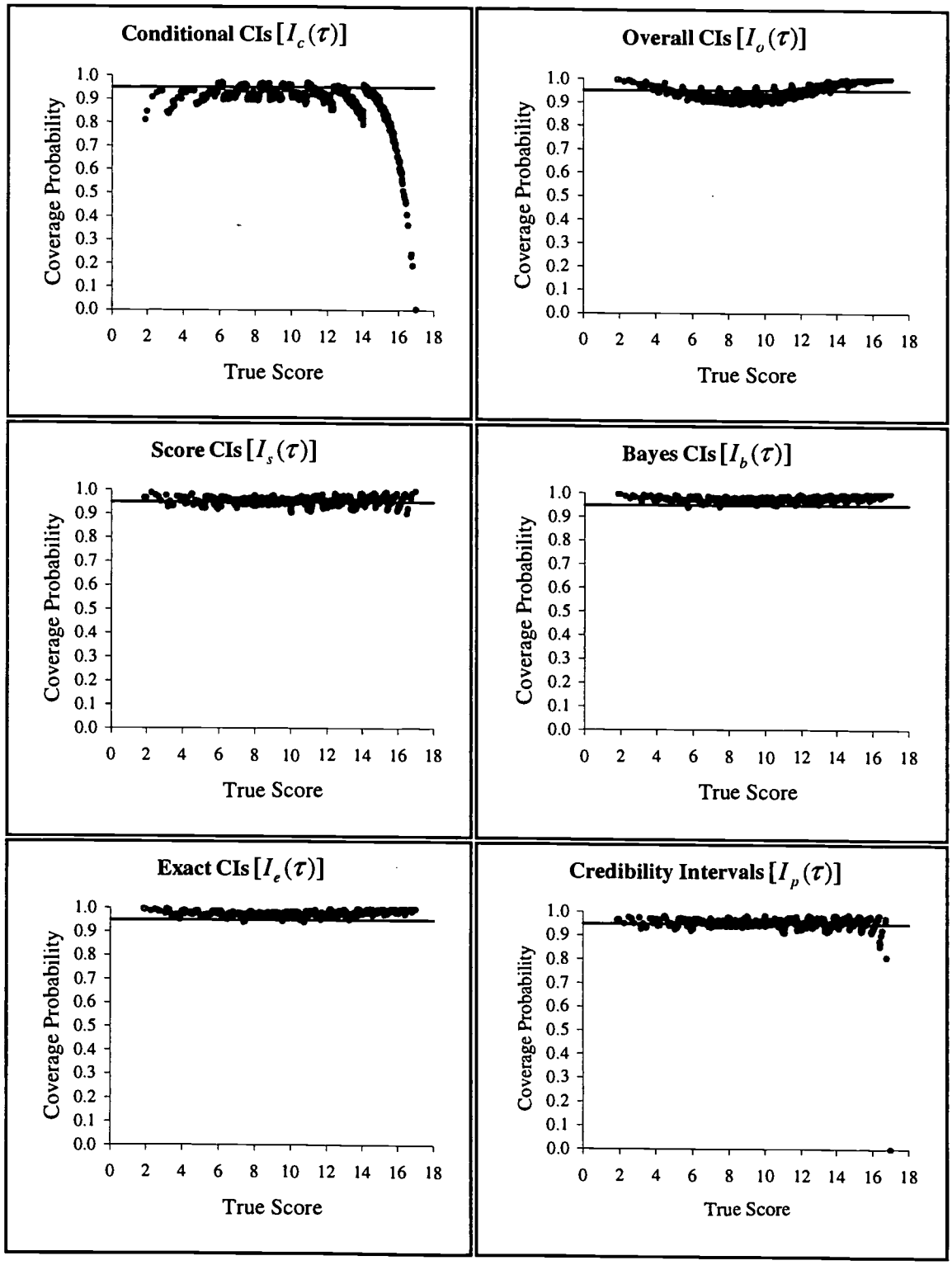


*FIGURE 20.* **Actual Coverage Probabilities of Nominal 95% Scale - Score Intervals Using Conditional Scale - Score SEMs $[I_c(\xi)]$ with $k = 17$**

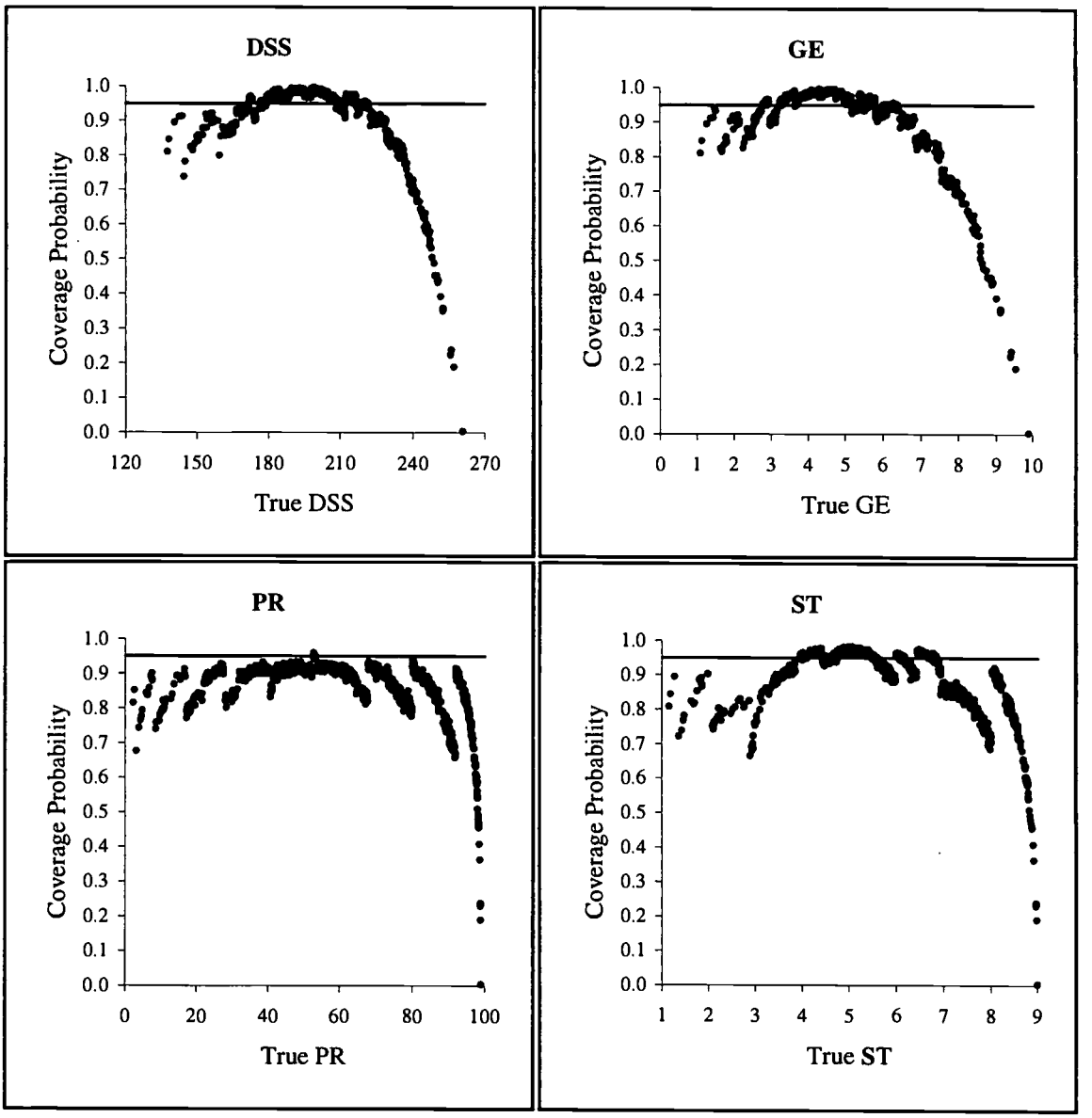


*FIGURE 21.* **Actual Coverage Probabilities of Nominal 95% Scale - Score Intervals Using Overall Scale - Score SEMs $[I_o(\xi)]$ with $k = 17$**

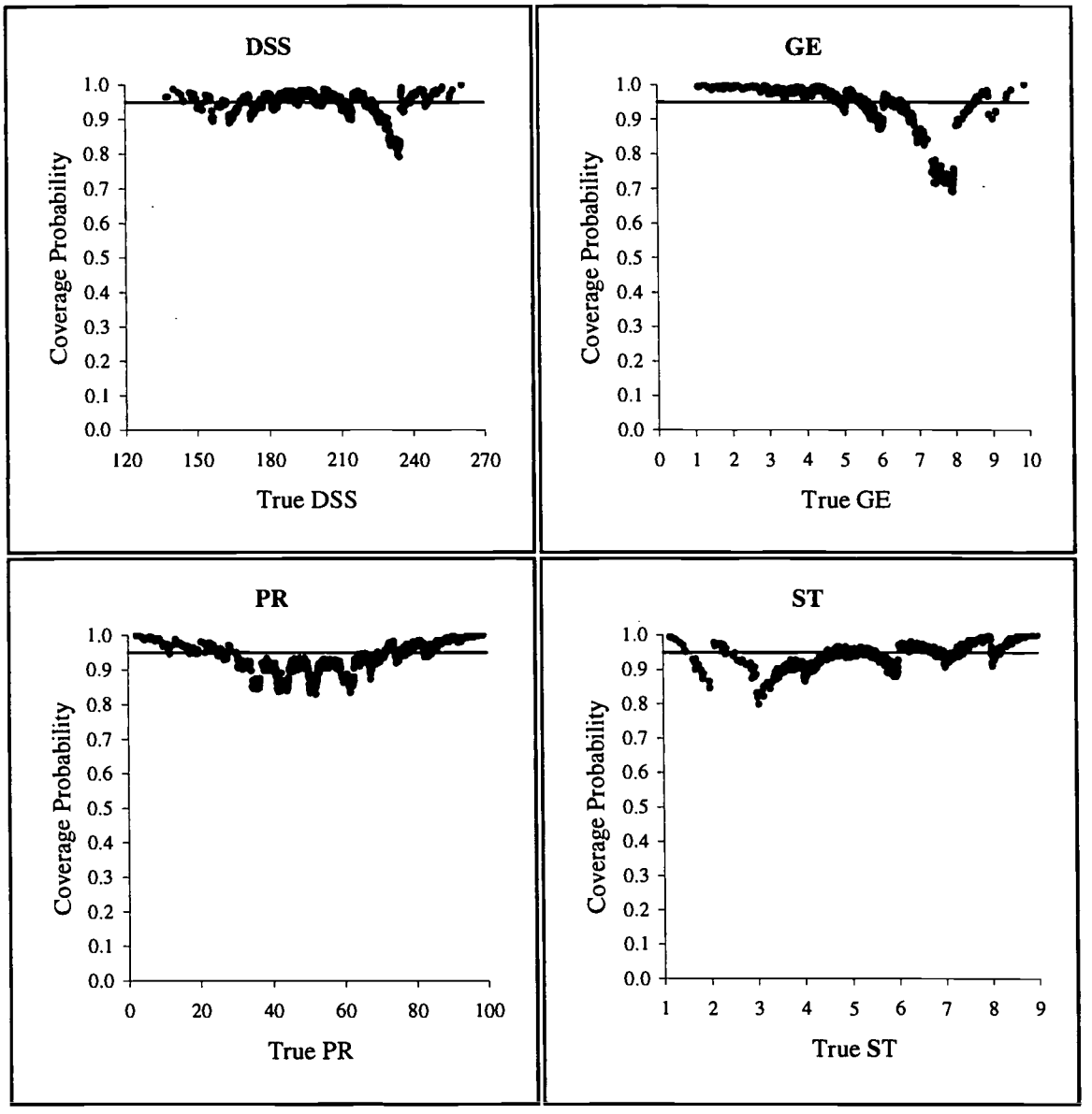


*FIGURE 22* . **Actual Coverage Probabilities of Nominal 68% Raw-Score Intervals with $k$ = 17**

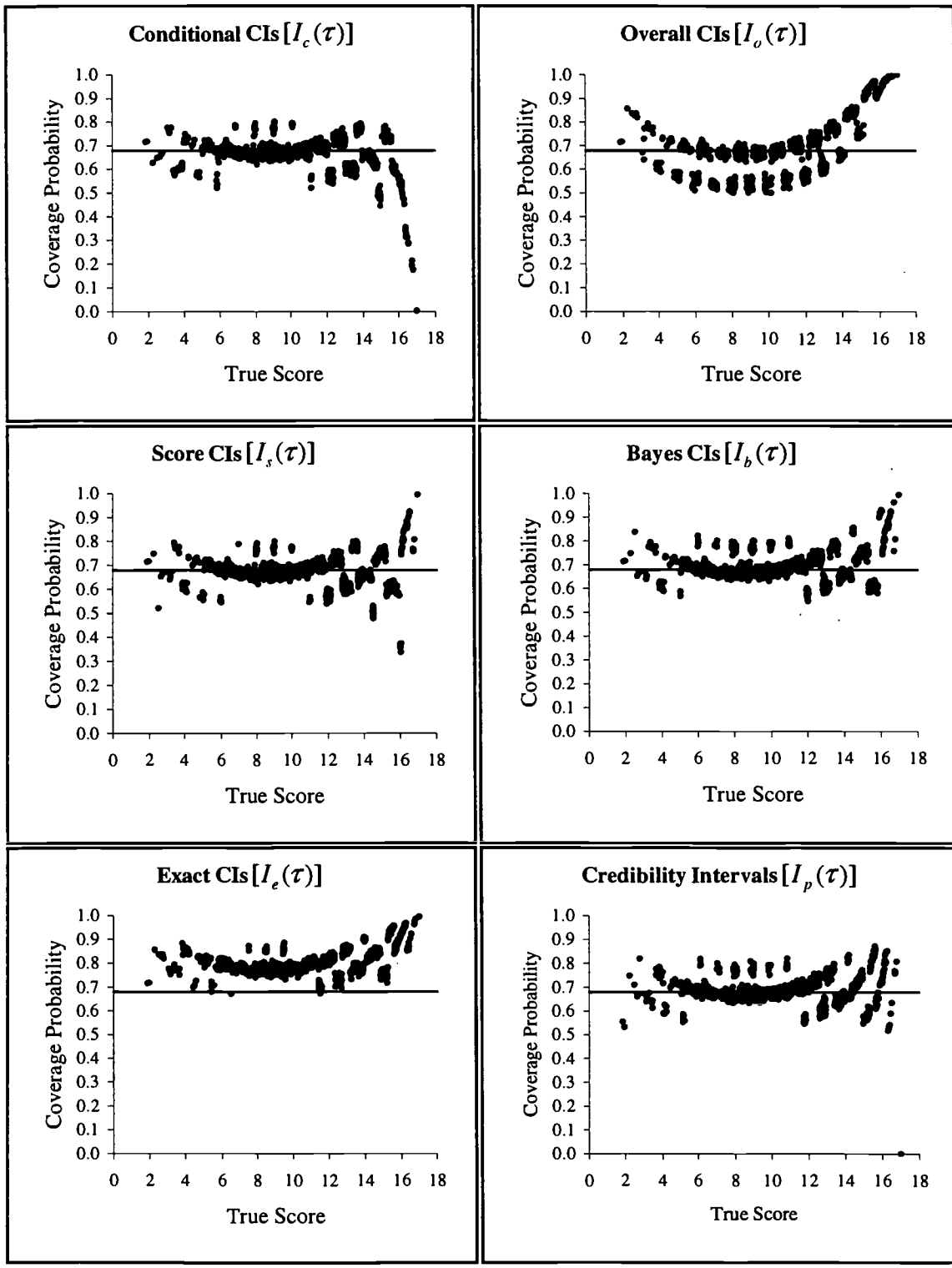


*FIGURE 23.* **Actual Coverage Probabilities of Nominal 68% Scale - Score Intervals Using Conditional Scale - Score SEMs $[I_c(\xi)]$ with $k = 17$**

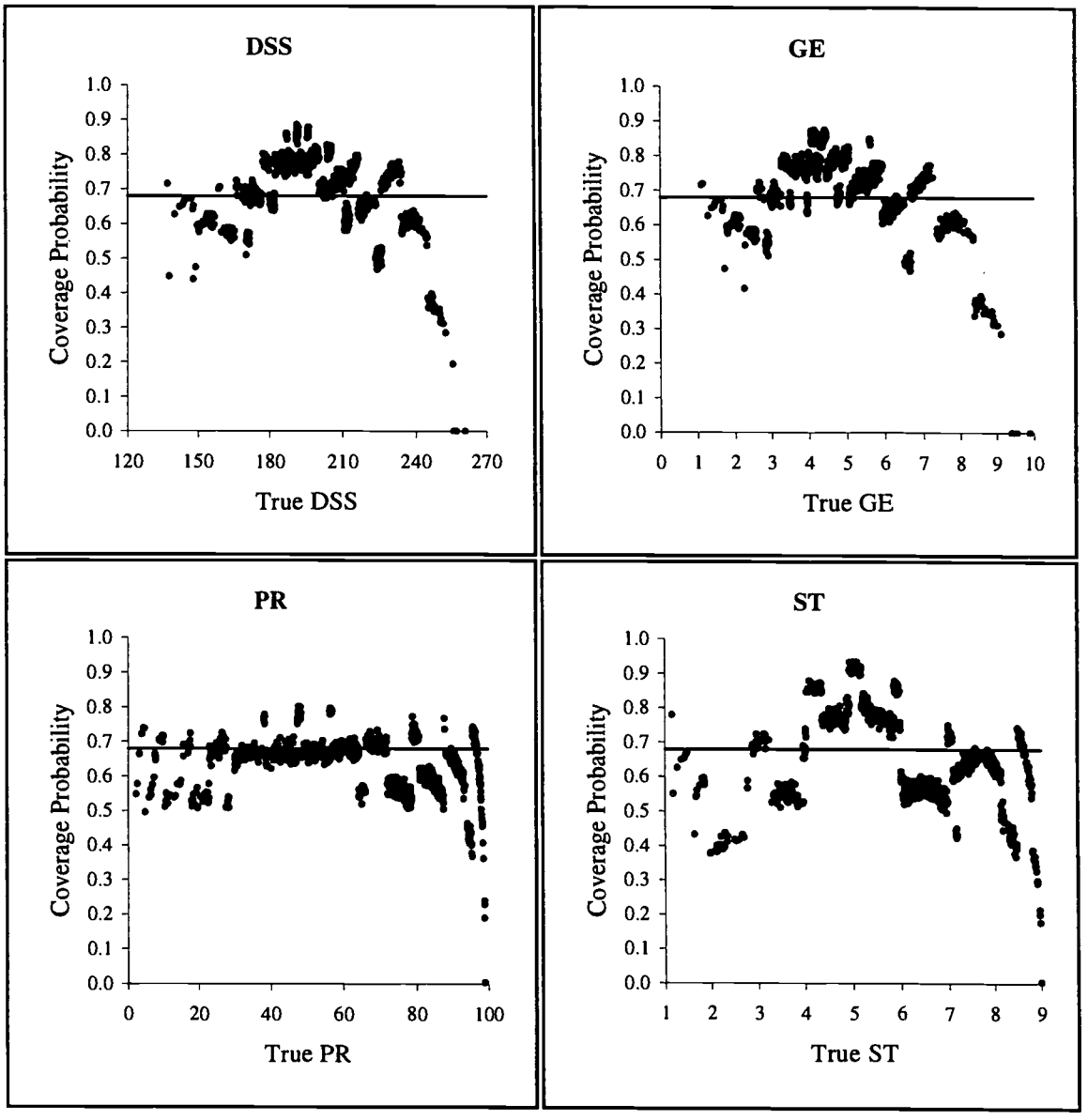


*FIGURE 24.* **Actual Coverage Probabilities of Nominal 68% Scale - Score Intervals Using Overall Scale - Score SEMs $[I_o(\xi)]$ with $k = 17$**

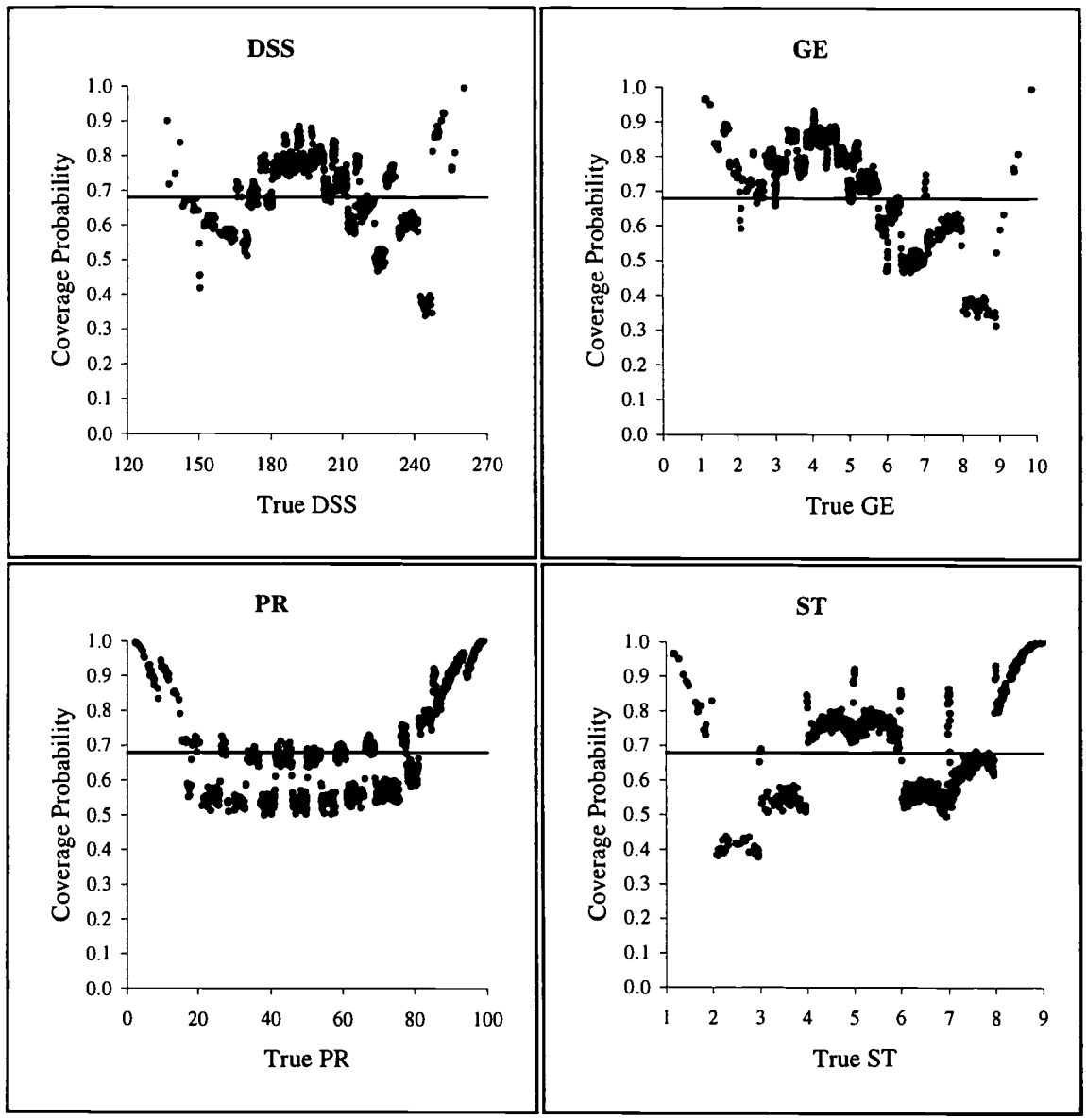


***FIGURE 25*. Actual Coverage Probabilities of Nominal 50% Raw-Score Intervals with $k = 17$**

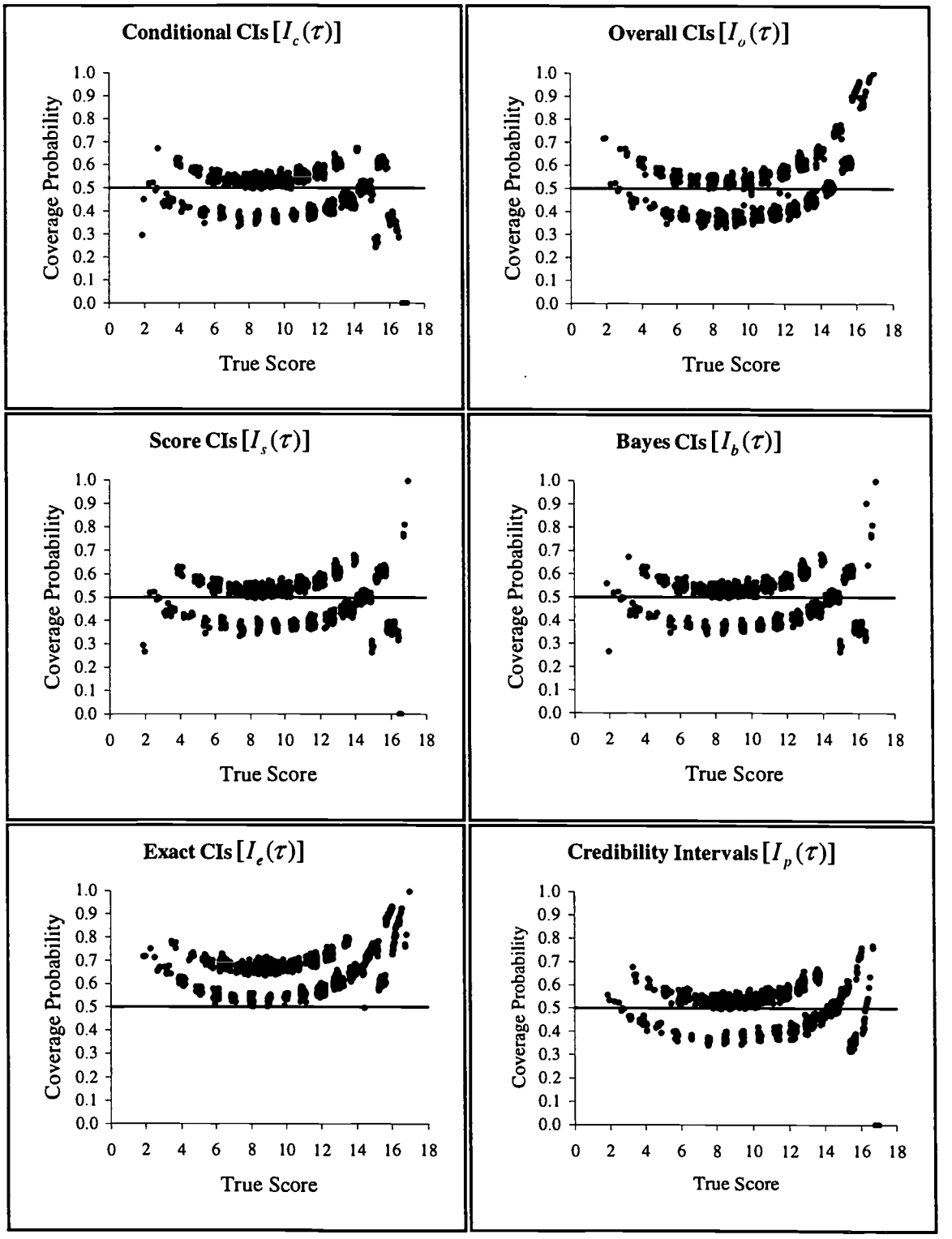


*FIGURE 26.* **Actual Coverage Probabilities of Nominal 50% Scale - Score Intervals Using Conditional Scale - Score SEMs $[I_c(\xi)]$ with $k = 17$**

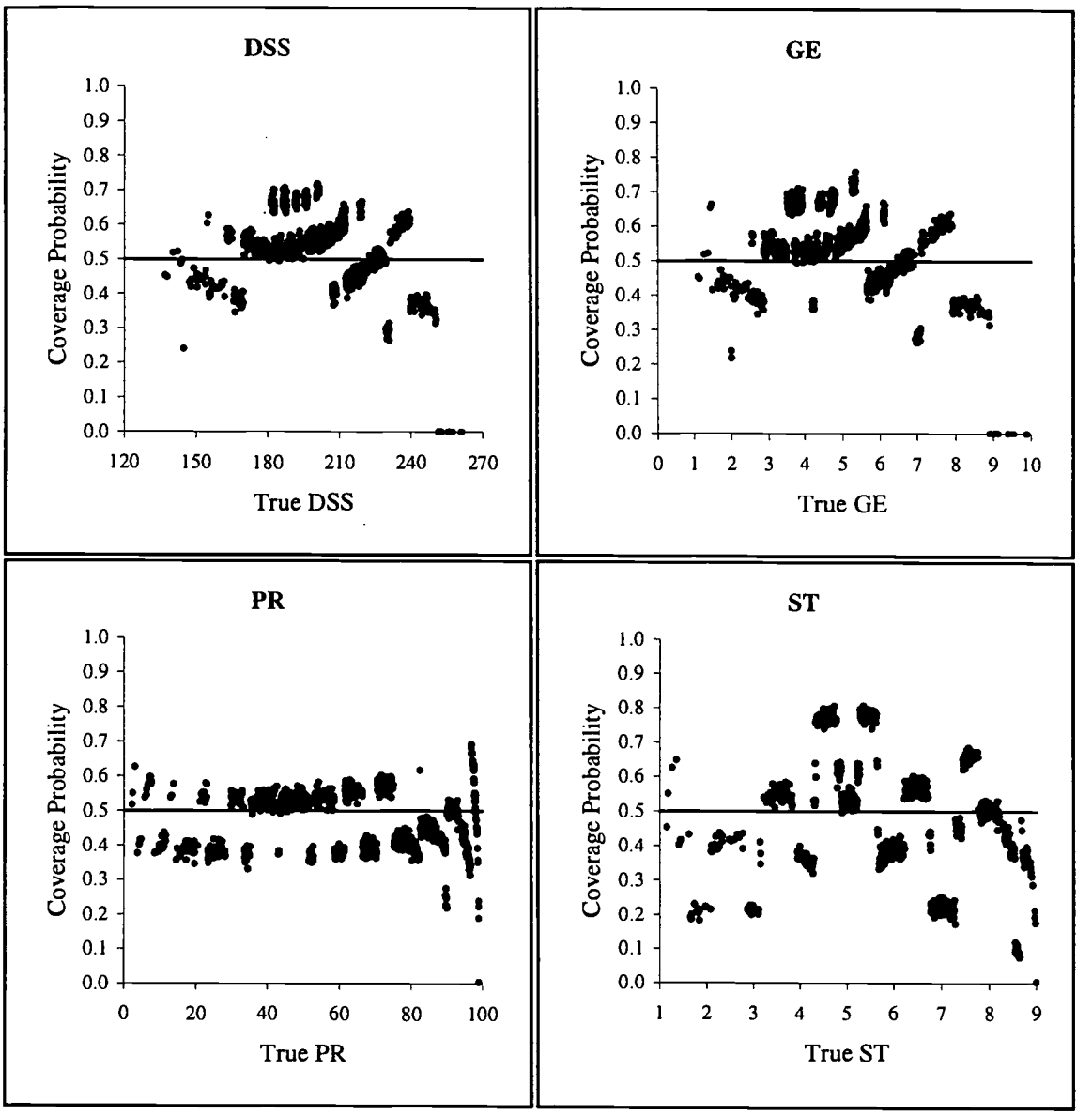
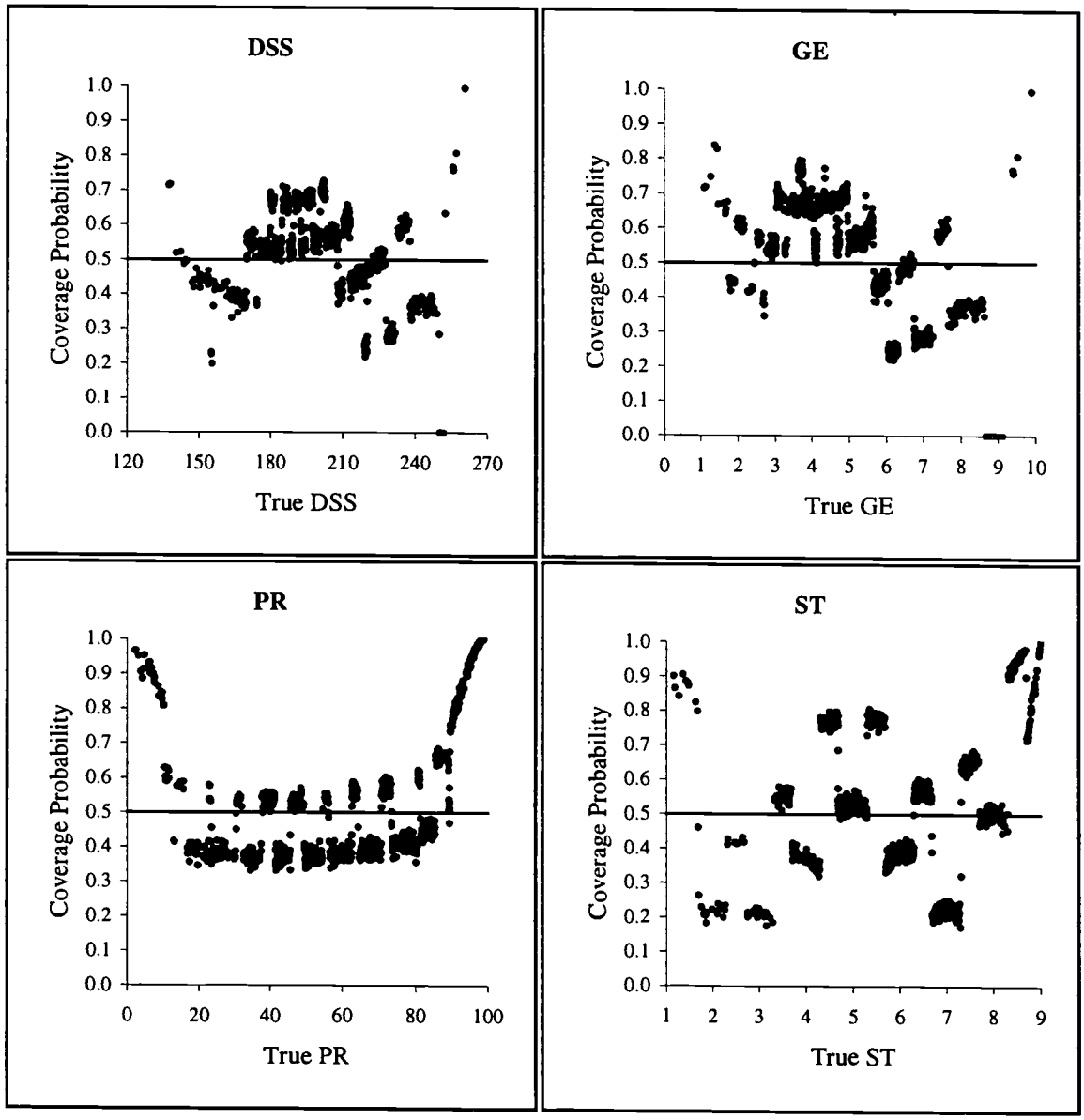


*FIGURE 27.* **Actual Coverage Probabilities of Nominal 50% Scale - Score Intervals Using Overall Scale - Score SEMs $[I_o(\xi)]$ with $k = 17$**




**U.S. Department of Education**
*Office of Educational Research and Improvement (OERI)*
*National Library of Education (NLE)*
*Educational Resources Information Center (ERIC)*


# NOTICE

## Reproduction Basis

This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

OBTAIN

SAMPLE

**ERIC_NO:** ED462418
**TITLE:** Data Sparseness and Online Pretest Item Calibration/Scaling Methods in CAT. *ACT Research Report* Series.
**AUTHOR:** Ban, Jae-Chun; Hanson, Bradley A.; Yi, Qing; Harris, Deborah J.
**PUBLICATION_DATE:** 2002
**ABSTRACT:** The purpose of this study was to compare and evaluate three online pretest item calibration/scaling methods in terms of item parameter recovery when the item responses to the pretest items in the pool would be sparse. The three methods considered were the marginal maximum likelihood estimate with one EM cycle (OEM) method, the marginal maximum likelihood estimate with multiple EM cycles (MEM) method, and Stocking's Method B. The three methods were evaluated using simulations of data from computerized adaptive tests (CAT). The MEM method produced the smallest average total error in recovering the 240 pretest item characteristic curves. Stocking's Method B yielded the second smallest average total error in parameter estimation. In terms of scale maintenance, the MEM method and Stocking's Method B performed well in keeping with the scale of the pretest items on the same scale as that of the true parameters. With the OEM method, the scale of the pretest item parameter estimates deviated from that of the true parameters. (Contains 1 figure, 4 tables, and 14 references.) (Author/SLD)
**MAJOR_DESCRIPTORS:** Adaptive Testing; Computer Assisted Testing; Online Systems; Pretests Posttests; Scaling;
**MINOR_DESCRIPTORS:** Data Analysis; Error of Measurement; Estimation (Mathematics); Maximum Likelihood Statistics; Simulation; Test Items;
**IDENTIFIERS:** *Calibration; EM Algorithm
PUBLICATION_TYPE: 142
**PAGE:** 23
**CLEARINGHOUSE_NO:** TM033682
**AVAILABILITY:** ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243-0168. Tel: 319-337-1028; Web site: http://www.act.org.
EDRS_PRICE: EDRS Price MF01/PC01 Plus Postage.
**INSTITUTION_NAME:** JXQ01640 _ American Coll. Testing Program, Iowa City, IA.
**REPORT_NO:** ACT-RR-2002-1
LEVEL: 1
**LANGUAGE:** English
**GEOGRAPHIC_SOURCE:** U.S.; Iowa
**ERIC_ISSUE:** RIEAUG2002